



Scaffold

Version 4.0

User's Manual

Release information

The following release information applies to this version of the *Scaffold User's Manual*. This document is applicable for Scaffold, Release 4.0 or greater, and is current until replaced.

<i>Document Version Number</i>	<i>Scaffold-4-UG4.3.0</i>
<i>Document Status</i>	<i>Released</i>
<i>Document Release Date</i>	<i>February 12, 2014</i>

Copyright

© 2014. Proteome Software, Inc., All rights reserved.

The information contained herein is proprietary and confidential and is the exclusive property of Proteome Software, Inc.. It may not be copied, disclosed, used, distributed, modified, or reproduced, in whole or in part, without the express written permission of Proteome Software, Inc.

Limit of Liability

Proteome Software, Inc.. has used their best effort in preparing this guide. Proteome Software, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this guide and specifically disclaims any implied warranties of merchantability or fitness for a particular purpose. Information in this document is subject to change without notice and does not represent a commitment on the part of Proteome Software, Inc. or any of its affiliates. The accuracy and completeness of the information contained herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every user.

The software described herein is furnished under a license agreement or a non-disclosure agreement. The software may be copied or used only in accordance with the terms of the agreement. It is against the law to copy the software on any medium except as specifically allowed in the license or the non-disclosure agreement.

Trademarks

The name *Proteome Software*, the Proteome Software logo, *Scaffold*, *Scaffold Q+*, *Scaffold Q+S*, and the Scaffold, Scaffold Q+, and Scaffold Q+S logos are trademarks or registered trademarks of Proteome Software, Inc. All other products and company names mentioned herein may be trademarks or registered trademarks of their respective owners.

Customer Support

Customer support is available to organizations that purchase *Scaffold*, *Scaffold Q+* or *Scaffold Q+S* and that have an annual support agreement. Contact Proteome Software at:

Proteome Software, Inc.

1340 SW Bertha Blvd, Suite 10

Portland, OR 97219

1-800-944-6027 (Toll Free)

1-503-245-4910 (Fax)

www.proteomesoftware.com

Table of Contents

Preface	7
Chapter 1: Getting Started with Scaffold	11
Chapter 2: Identifying Proteins with Scaffold	23
Chapter 3: Loading Data in Scaffold	37
Chapter 4: The Scaffold Window	61
Chapter 5: The Load Data View	111
Chapter 6: The Samples View	121
Chapter 7: Quantify View	143
Chapter 8: Protein Grouping and Clustering	151
Chapter 9: Quantitative Methods and tests	163
Chapter 10: Precursor Intensity Quantitation	179
Chapter 11: Reports.....	189
Appendix.....	205

Preface

Welcome to the *Scaffold User's Manual*. The purpose of the *Scaffold User's Manual* is to answer Users' questions and guide them through the procedures necessary for using Scaffold efficiently and effectively.

Using the manual

The *Scaffold User's Manual* is easy to use. The User can simply look up the topic that he/she needs in the table of contents or the index. Later, in this Preface, a brief discussion of each chapter is provided to further assist the User in locating the information that he/she needs.

Special information about the manual

The *Scaffold User's Manual* has a dual purpose design. It can be distributed electronically and then printed on an as-needed basis, or it can be viewed on-line in its fully interactive capacity. If the User prints the document, for best results, it is recommended that he/she prints it on a duplex printer; however, single-sided printing will also work. If the User views the document on-line, a standard set of bookmarks appears in a frame on the left side of the document window for navigation through the document. For better viewing, decreasing the size of the bookmark frame and using the magnification box to adjust the magnification of the document will help the User in setting his/her viewing preference.



If the User decides to print the document using a single-sided printer, he/she might see a single blank page at the end of some chapters. This blank page has been added solely to ensure that the next chapter begins on an odd-numbered page. This blank page in no way indicates that the book is missing information.

Conventions used in the manual

The *Scaffold User's Manual* uses the following conventions:

- Information that can vary in a command—variable information—is indicated by alphanumeric characters enclosed in angle brackets; for example, <ProteinName>.
- A new term, or term that must be emphasized for clarity of procedures, is *italicized*.
- Page numbering is “on-line friendly.” Pages are numbered from 1 to x, *starting with the cover* and ending on the last page of the index.
- This manual is intended for both print and on-line viewing.

- Although numbering begins on the cover page, this number is not visible on the cover page or front matter pages. Page numbers are visible beginning with the first page of the Table of Contents.
- If information appears in [blue](#), it is a hyperlink. Table of Contents and Index entries are also hyperlinks. Click the hyperlink to advance to the referenced information.
- The example experiments data and databases available for download in zip format on Proteome Software's website at the following link: www.proteomesoftware.com/products/demo-data/#scaffold/ is used as the basis for most screen captures, examples, and data manipulations that are shown in the manual.

Assumptions for the manual

The *Scaffold User's Manual* assumes that:

- You are familiar with Windows operating systems, and basic Windows navigational elements, content formatting and layout tools.
- You have the appropriate licensing to run Scaffold.
- You have downloaded one of the three example experiments, available at www.proteomesoftware.com/products/demo-data/#scaffold/



- Choose SEQUEST or Mascot Samples and download the appropriate zip file.
- When the download has finished, move the file to the desired location on your hard drive and unzip it. You'll have a folder entitled *scaffold_tutorial* containing:
 - Sample search engine files results and related databases to be used in the loading example described in [Chapter 3, "Loading Data in Scaffold," on page 37](#).

The databases provided in the downloads are subset databases that will allow the tutorial searches to complete in a relatively short time. They do not necessarily generate complete protein identification results.



Further example exercises with guided explanations detailing different aspects of the way Scaffold can be used are available at <http://www.proteomesoftware.com/products/scaffold/> under the section Scaffold Tutorials in the left side menu.

Organization of the manual

In addition to this Preface, the *Scaffold User's Manual* contains the following chapters:

- [Chapter 1, "Getting Started with Scaffold," on page 11](#), which explains the tiered license structure for the Scaffold application suite. It also explains how to start Scaffold and also

details the different types of data that can be analyzed in Scaffold, Scaffold Q+ and Scaffold Q+S.

- [Chapter 2, “Identifying Proteins with Scaffold,” on page 23](#), which introduces the different views available in Scaffold to help mass spectrometrists and medical researchers confidently identify proteins in biological samples..
- [Chapter 3, “How Scaffold Structures Data,” on page 35](#), which introduces the User to the way Scaffold thinks about an experiment, the type of data it loads and what can be done to have an in depth look at the search results loaded in the program.
- [Chapter 3, “Loading Data in Scaffold,” on page 37](#), which guides the first time User step by step through the Scaffold Loading Wizard with an example using real data.
- [Chapter 4, “The Scaffold Window,” on page 61](#), which provides a detailed description of the main Scaffold window with all the tools it includes.
- [Chapter 5, “The Load Data View,” on page 111](#), which includes information about the search data loaded in the current Scaffold experiment with all the loading tools it includes
- [Chapter 6, “The Samples View,” on page 121](#), which provides a description of the functionality of the view, of the Samples table with all the available tools for filtering and searching specific proteins present in the list.
- [Chapter 7, “Quantify View,” on page 143](#), which provides graphical tools to help the User visualize experiments and draw conclusions about the quantitative relationships demonstrated in the data.
- [Chapter 8, “Protein Grouping and Clustering,” on page 151](#), which provides a detailed explanation of the grouping and clustering algorithms included in Scaffold
- [Chapter 9, “Quantitative Methods and tests,” on page 163](#), which provides a description of the different quantitative statistics and quantitative statistical tests available in Scaffold.
- [Chapter 10, “Precursor Intensity Quantitation,” on page 179](#), which provides a comprehensive description of how Scaffold treats and computes precursor intensity quantitation.
- [Chapter 11, “Reports,” on page 189](#), which includes a description of the various exports available.

Preface

Chapter 1

Getting Started with Scaffold

Scaffold is a software tool designed to help scientists identify and analyze proteins in biological samples. Using output files from MS/MS search engines, Scaffold validates, organizes, and interprets mass spectrometry data, allowing the User to more easily manage large amounts of data, compare samples, and search for protein modifications.

The Scaffold Viewer is a free, read-only version of Scaffold available Online for download. It facilitates the sharing of Scaffold analysis results among collaborators.

This chapter covers the following topics:

- [“Initial requirements” on page 12](#), which describes the minimum requirements for installing and running Scaffold.
- [“Scaffold Tiered Licensing” on page 13](#), which explains the type licenses available for activating the program.
- [“Scaffold Viewer” on page 18](#), free download.
- [“ScaffoldBatch” on page 19](#), which loads and analyzes the same data that Scaffold does, but in a batch rather than in an interactive environment.
- [“How Scaffold structures data” on page 20](#), which describes the format of the files that are compatible with Scaffold.

Initial requirements

Before installing and running Scaffold the User needs to make sure that:

1. The computer system where Scaffold is going to be installed and its network must have access to directories containing:
 - Search engine output files for the samples that need to be analyzed
 - the FASTA database(s) used when those files were run.
2. Check the following document for general system requirement:
[System_requirements.pdf](#)
3. Check the following document for input files supported in Scaffold:
[File_compatibility_matrix.pdf](#)
4. Have a license key to run Scaffold, see [Scaffold Tiered Licensing](#).

Once installed, to run Scaffold the User needs to:

1. Either select the menu option **File > New** or click the **Add BioSample** button in the Load Data View, to open the Load Wizard. The Wizard helps the User go through the process of loading and analyzing data.

The first-time User when Scaffold initially opens needs to click the Run Demo button in the Welcome to Scaffold box. Then, open one of the previously saved tutorial files to start playing around with an existing experiment. Guided tutorials are also available at the following link: proteome-software.wikispaces.com/Tutorials

Scaffold Tiered Licensing

The Scaffold suite of applications consists of the core Scaffold product, Scaffold Q+, and Scaffold Q+S. The core Scaffold product is the basis for all installations. The licensing key that Proteome Software provides determines whether the User has access to just the features and functions of the core Scaffold product, or the features and functions of Scaffold Q+ or Scaffold Q+S.



Users who purchased a license for Scaffold Q+S, then also have access to all the features and functions for both Scaffold and Scaffold Q+.

Application	Description
Scaffold	Visualize and validate MS/MS proteomics experiments.
Scaffold Q+	Calculate and display relative protein expression levels in a sample determined by tandem mass spectrometry of iTRAQ- or TMT-labeled proteins.
Scaffold Q+S	Calculate and display relative protein expression levels in a sample determined by tandem mass spectrometry of stable isotopically-labeled (for example, SILAC) proteins.

After Scaffold has been installed on a computer, a shortcut icon for the application is placed on the desktop. An option is also available from the Start menu. The User can double-click the desktop icon to launch Scaffold, or select the option from the Start menu **Start > All Programs > Scaffold 4> Scaffold 4**.

Figure 1-1: Scaffold desktop icon

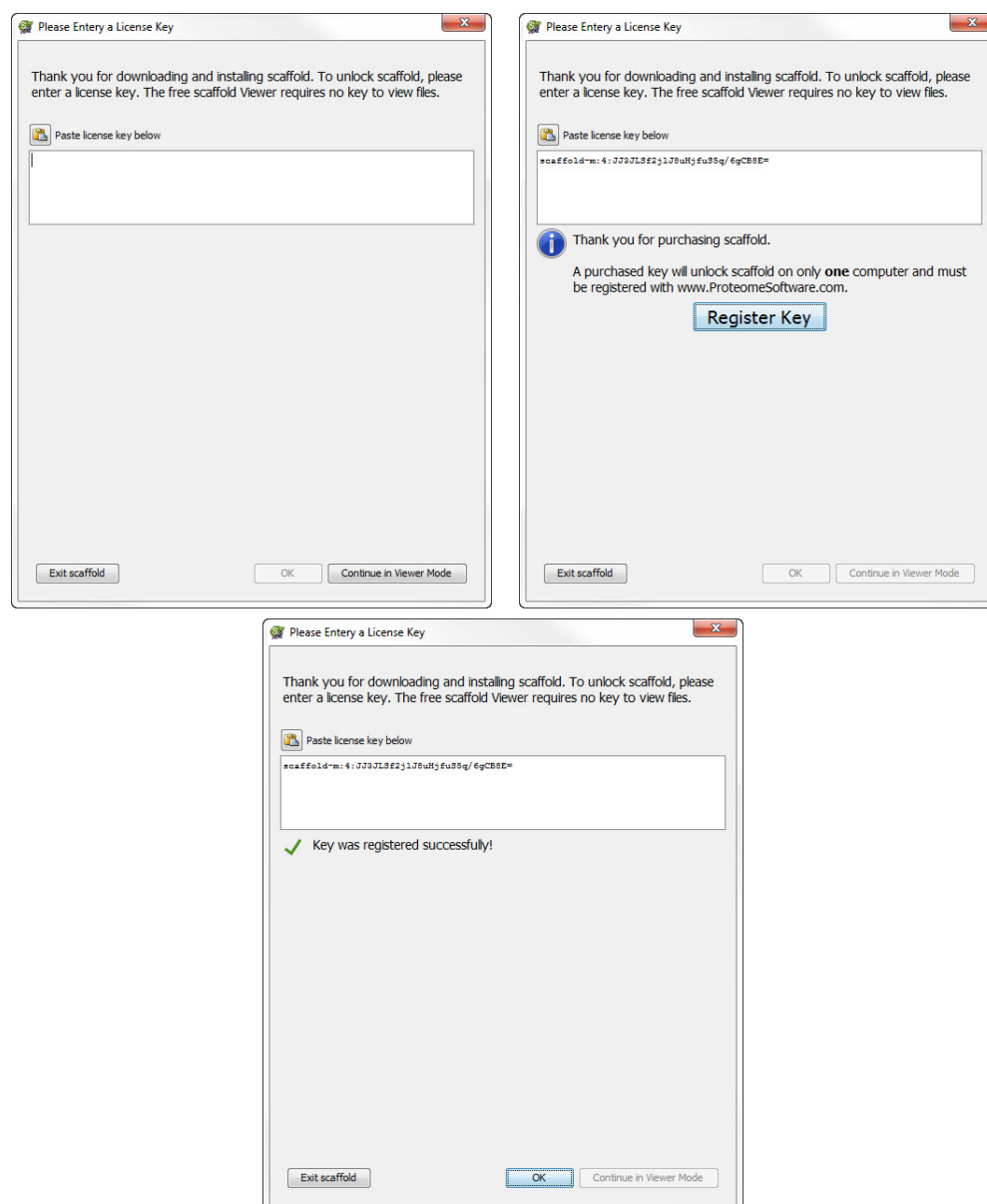


The *first* time the User opens Scaffold after installing it, the Enter License Key dialog box opens in the Scaffold main window.

Chapter 1

Getting Started with Scaffold

Figure 1-2: Scaffold License Key messages



There are two kinds of keys:

- **Evaluation key**—An Evaluation key is valid for two weeks. The User can obtain a free evaluation key for any of the Scaffold applications at www.proteomesoftware.com. The User can use this key on an unlimited number of computers.

- **Time-Based License key**—a Time-Based License key allows the User to access all features of the software permanently. It only allows upgrades within a certain time limit, however. The time tracks the length of the support contract. Once expired, Scaffold continues to work beyond the key expiration date, but no new upgrades are allowed unless the support contract is renewed. The user must contact sales@proteomesoftware.com to purchase the appropriate key. A Time-Based License key is valid for only a single computer. If the user moves the Scaffold installation to a different computer, he/she should contact sales@proteomesoftware.com to transfer the key at no charge.

After the User enters the key and presses OK, the Key dialog box closes and a Scaffold Welcome message opens. The Welcome message and the title bar for the Scaffold main window indicate the application to which the User has access—Scaffold, Scaffold Q+ or Scaffold Q+S.



If the User is using an evaluation copy of Scaffold, then an Evaluation message opens, indicating the number of days t left in the evaluation period. The User must click OK to close this message and then the Scaffold Welcome message opens.

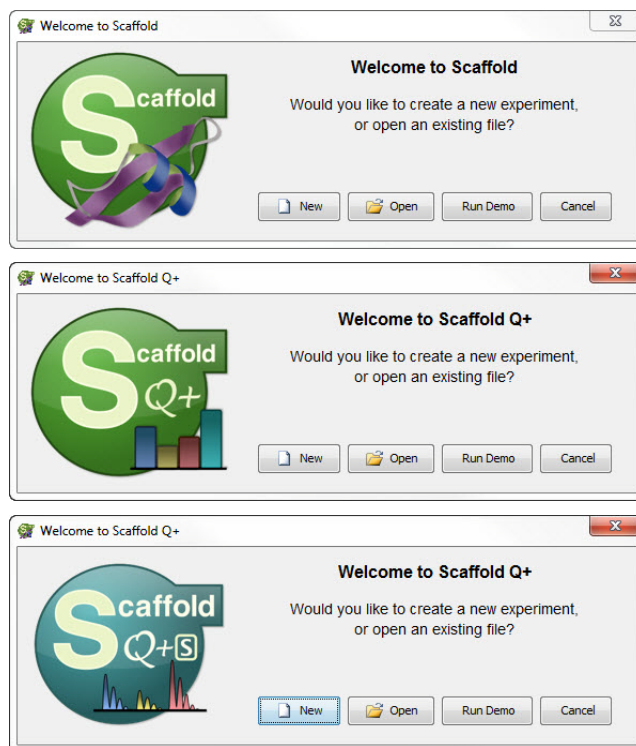
Figure 1-3: Welcome Window Scaffold version <Version #> indicating access to Scaffold, Scaffold Q+ or Scaffold Q+S



Chapter 1

Getting Started with Scaffold

Figure 1-4: Scaffold main Welcome window indicating access to Scaffold, Scaffold Q+ or Scaffold Q+S



From this window, the User can create a new experiment, open an existing experiment (*.SF3 file), or work with the demonstration data that is provided with all Scaffold installations.

Upgrading Scaffold to Scaffold Q+ or Scaffold Q+S

When the User is running a core copy of Scaffold and would like to upgrade to Scaffold Q+ or Scaffold Q+S he can do so by contacting our sales department at sales@proteomesoftware.com. When the purchase is finalized the sales department will send the User an upgraded license key that will unlock the Scaffold Q+ or Scaffold Q+S features.

To input the new upgraded key the User should follow the following instructions:

1. Open the current copy of Scaffold installed on the computer by double clicking on the Scaffold icon found on the desktop or selecting the Scaffold application from the start up menu.
2. Make sure you are connected to the Internet
3. If no upgrades are available click continue on the first “Welcome to Scaffold” dialog.

4. If a warning appears suggesting to upgrade Scaffold do so and open Scaffold after the upgrade.
5. When the second “Welcome to Scaffold” dialog appears click cancel.
6. Go to the **Help** menu and select the option **Upgrade License Key...**
7. When the **Overwrite** dialog appears click Yes.
8. The **Fully licensed Scaffold** dialog opens showing the information related to the current license.

Figure 1-5: Fully Licensed Scaffold dialog



9. Click Enter New Key and the **Please Enter License Key** dialog opens, see [Figure 1-2](#).
10. Copy and paste the license key. After verification of the key the Register Key button appears, click it.
11. If the key is valid the message “Key was registered successfully!” appears, click OK Scaffold is ready to go.
12. If the key is not valid for whatever reason contact back sales@proteomesoftware.com.

Renewing time based license key

Time based license keys have time limits connected to their validity. When a time based key expires Scaffold still works but upgrades are not allowed until the support contract is renewed. The status of the Scaffold license key can be checked in the **About Scaffold...** dialog the User opens selecting **Help > About Scaffold** command from the main menu.

If the key is expired and the User wants to upgrade Scaffold, clicking the Renew button in the dialog opens the **Key reset Request** page on the Proteome Software website. The User needs to fill in the request and a sales representative will promptly contact him/her providing further information.

Scaffold Viewer

When a licensed copy of Scaffold is installed on a computer only one full copy at a time can run on the system. On the other hand the User can open multiple copies of the Viewer at the same time. Scaffold Viewer can open and read any *.SF3 file created by Scaffold.

The Viewer is free, and Users may install it on as many computers as they wish. If a User analyzes data with Scaffold, he/she can give a copy of the Viewer to all his/her collaborators so that they can view the User's data.

The Viewer performs most of the functions included in a full Scaffold copy. However, it cannot load any of the search results files and it can neither analyze data nor run X! Tandem.

With the Viewer the User (or his/her collaborators) can look at the data in the same ways as with Scaffold: by samples, proteins, peptides or spectra. The User can filter the results by protein probability, peptide probability and the number of matching peptides or FDR values. The User can change the names of the BioSamples, the MS Samples, and the proteins. The Viewer User can also validate the peptide/spectrum matches.

ScaffoldBatch

ScaffoldBatch is a batch version of Scaffold. It can load and analyze the same data that Scaffold does, but in a batch rather than in an interactive environment. Batch mode means that ScaffoldBatch can be run on the command line or called from a batch script. The intended use is for organizations that want to integrate Scaffold into a Proteomics pipeline.

ScaffoldBatch can be used as one component of an automated Proteomics work flow. ScaffoldBatch is an extended version of Scaffold.

When the User installs ScaffoldBatch, a copy of the interactive version of Scaffold is automatically installed as well. Like Scaffold, ScaffoldBatch is locked to one computer by a software key. As a command line batch program, ScaffoldBatch is intended to be called from a batch script. In the Microsoft world this might be a *.BAT file. In the Linux world this might be a *.SH file.

ScaffoldBatch is driven by an XML file (*.SCAFML) that specifies all the needed operations to create a *.SF3 Scaffold file experiment. For more technical detailed information about how to install and run ScaffoldBatch the User can consult the ScaffoldBatch manual at: www.proteomesoftware.com/pdf/scaffold_batch_users_guide.pdf

How Scaffold structures data

Scaffold stores all the data related to an experiment in one single file. Each experiment file (*.SF3) can hold up to 600,000 spectra and associated data. The User can create, name, and save as many experiment files as disk space permits, but only one at a time can be opened with full Scaffold capabilities. Multiple experiments can be opened in the Viewer mode.

Experimenters frequently categorize biological samples in larger groups to compare, for example, diseased with control, treated with control, day1 with day2, pregnant with not pregnant. To capture this, Scaffold associates a sample category with each biological sample.

Data associated with a biological sample (abbreviated in Scaffold as BioSample) comes from a sample taken by a doctor, medical researcher, or biologist, such as a drop of blood or biopsy from a patient, or a tissue sample from a model organism or cell line. Using such techniques as 2D gels or liquid chromatography, proteins or peptides from these biosamples are then separated from each other. Each resulting individual band, spot, or LC fraction then processed by a mass spectrometer is one mass spectrometry sample (abbreviated in Scaffold as MS sample).

One BioSample is therefore typically made up of more than one MS sample — sometimes many more.

Scaffold can also process data from MuDPIT experiments, in which case the analysis combines peptides from all fractions into one MS sample for protein identification.

Data Loading

Scaffold imports data generated from a large variety of search engines like Mascot, SEQUEST Spectrum Mill, OMSSA, Phenyx, X!Tandem, MaxQuant. It also supports those search engines that can export the search results in the mzIdentML format. All type of search data can be freely included in one experiment. Each SEQUEST folder is imported as one “file,” as is each Mascot or X! Tandem file.

Importing files requires access rights from the computer where Scaffold is installed to the location where the search results data files reside.

The loading data phase is also where BioSamples are defined and as part of the import process, the User can specify all the files he/she wishes to include in a BioSample, which can then be named and categorized.

The User can load data files in Scaffold either activating [The Loading Wizard](#) by selecting New from the [Main menu commands](#) or using the selections available in the [The Load Data View](#). The following document, published on the Proteome Software website, provides detailed information on search engine data files compatible with Scaffold, Scaffold Q+ or Scaffold :[File_compatibility_matrix.pdf](#)

Quantitative Data

Scaffold Q+ and Q+ S are Proteome Software's labeling quantitation software packages.

- Q+ loads iTRAQ (Applied Biosystems) and Tandem Mass Tagged (TMT, Thermo Scientific) labeled data.
- Q+S can also load stable isotope labeled samples.

If the User has purchased Scaffold Q+ or Scaffold Q+S, he/she will use Scaffold's file importing wizard to load the search results of the labeled data.

Quantitative Data File compatibility

Please, check the file compatibility matrix at the following link:

[File_compatibility_matrix.pdf](#)

Characterizing data

The data imported in Scaffold are the results of a previous search against a particular FASTA protein database, using a particular search engine (SEQUEST, Mascot, X! Tandem or others), on a particular set of data.

When the User imports these data, Scaffold needs to know certain characteristics of the specific search, so the User in the loading phase will be asked to:

- [Specify the Database](#)
- Specify the parameters used for the search

Specify the Database

As part of the loading process, the User needs to specify the particular FASTA database that the initial search engine used to obtain the results. This database must also be stored on a location accessible to the system where Scaffold is installed. It is important to specify the correct database. Without this information, Scaffold cannot display the full sequence of amino acids in a peptide, nor, therefore, the sequence coverage.

All search engines like for example SEQUEST, Mascot, or X! Tandem store the name of the database they use with their results. If the User is uncertain of the database used, he/she can use a text editor to search the search engine output files for the database name. It's possible, though, that the correct database resides on a local network under a different name.

It is best to use the same database for all search results loaded into a specific Scaffold experiment. This permits Scaffold to accurately align proteins found in different samples.

Chapter 1
Getting Started with Scaffold

Chapter 2

Identifying Proteins with Scaffold

Scaffold is a tool designed with the aim of helping mass spectrometrists and medical researchers confidently identify proteins in biological samples. Using output data from most of the current search engines available like: SEQUEST®, Mascot®, MaxQuant, X! Tandem and many others, Scaffold validates, organizes, and interprets mass spectrometry data, so that a User can easily manage large amounts of data, compare samples, and search for protein modifications.

Scaffold makes it easier to search data repeatedly, using additional methods to find results that might otherwise be missed. For example, it enables the user to export unidentified spectra, which can then be searched against larger databases to find additional proteins.

Alternatively, Scaffold can export a new FASTA database consisting only of those proteins found in the loaded BioSamples to allow searching of unidentified spectra against the subset database using different parameters — for example, specifying other variable modifications.

Whether the aim of the user is broadening or deepening a search, Scaffold can then re-import the new data and bring to bear its tools for compiling, comparing, and analyzing the results.

This chapter covers the following topics:

- [“Scaffold Flexible Workflow” on page 24,](#) which provides a brief description of possible work-flows to improve the analysis of the data sets loaded in a Scaffold experiment.
- [“Increased Confidence Using Peptide and Protein Validation Algorithms” on page 26,](#) which describes the statistical validation methods used in Scaffold.
- [“Scaffold Views” on page 29,](#) which provides an overview of the different structural views available in the Scaffold window.

Scaffold Flexible Workflow

Scaffold supplements spectra search engines; it does not replace them. The user continues to run the output of his/her mass spectrometry experiments through SEQUEST, Mascot, MaxQuant, X!Tandem or whatever other search engine compatible with Scaffold, as usual. Results are then imported into Scaffold.



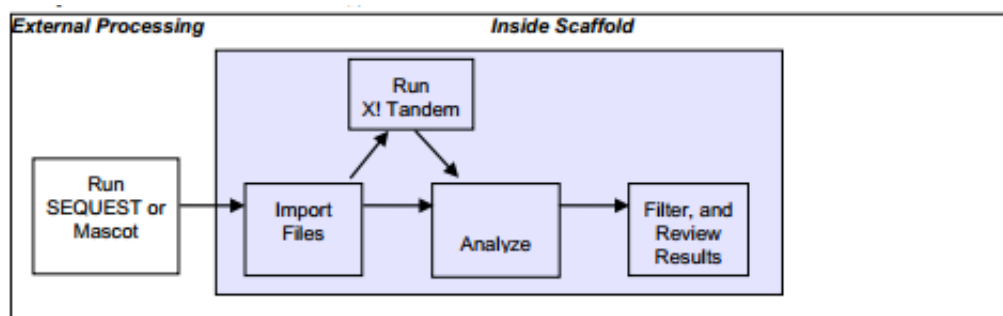
- For more information about compatibility with Scaffold check the following document: http://www.proteomesoftware.com/pdf/file_compatibility_matrix.pdf
- Scaffold comes bundled with X!Tandem. To increase identification confidence, the user can run the bundled version of X!Tandem on data previously analyzed by other search engines

Simple Workflow including Scaffold

Scaffold uses various scientifically validated, probabilistic methods to evaluate and analyze the imported data displaying its results in the Samples and Proteins views, for more information see references listed in the [Algorithms References](#) appendix. Once the data is loaded and analyzed by Scaffold, results are saved in special Scaffold files that bear the extension *.SF3. The Scaffold files so created can be closed and reopened again at a later time either through a full or a viewer version of Scaffold, see [Scaffold Viewer](#).

When a licensed copy of Scaffold is installed on a computer only one full copy at a time can run on the system. On the other hand the User can open multiple copies of the Viewer at the same time. Scaffold Viewer can open and read any *.SF3 file created by Scaffold.

Figure 2-1: Data Analysis Workflow including Scaffold



Broadening the Search

For further investigations the User can export the results to an Excel spreadsheet, or export the unidentified spectra in a format that allows further searching by SEQUEST or Mascot or MaxQuant or by any other compatible search engine. In this way, the original searches can be repeated with different parameters, or the results can be searched using another search

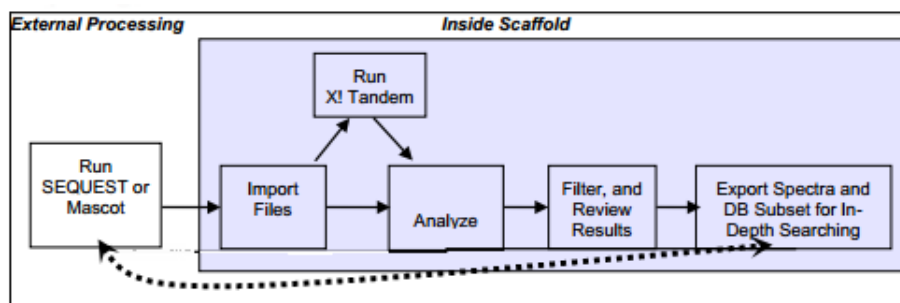
engine, or against a different database. Importing the new resulting output data into Scaffold, augments the initial Scaffold experiment.

Deepening the Search

Deepening the Search

If, rather than casting a wider net, a deeper look is preferred, the User can export the unmatched spectra and a subset database consisting only of those proteins already found in the BioSamples included in the Scaffold experiment. The spectra can then be searched against this subset database, allowing a relatively fast search of these known proteins using variations on search parameters or looking for additional modifications.

Figure 2-2: Deepening the search



Increased Confidence Using Peptide and Protein Validation Algorithms

For validating peptide and protein identifications Scaffold uses three different scoring systems:

- [LFDR-based scoring system](#) for peptide validation - Developed in Scaffold 4 it is particularly effective for high mass accuracy data (including data acquired on QExactive instruments)
- [PeptideProphet](#) scoring system - Bayesian statistical algorithm developed by the Institute for Systems Biology and available in Scaffold since its first version.
- [ProteinProphet](#) scoring algorithm - Bayesian statistical algorithm developed by the Institute for Systems Biology and available in Scaffold since its first version.

Implementations of the last two algorithms have been widely distributed under the names PeptideProphet™ and ProteinProphet®. Scaffold uses an independent implementation of these algorithms, for more information see [Algorithms References](#), [Keller \(2002\)](#) and [Nesvizhskii \(2003\)](#).

LFDR-based scoring system

In this method, peptide identifications are validated by discriminant scoring using a Naïve Bayes classifier generated through iterative rounds of training and validation to optimize training data set choices. Peptide probabilities are assessed using a Bayesian approach to local FDR (LFDR) estimation. Rather than just using mass accuracy as a term in discriminant score training, peptide probabilities are modified by likelihoods calculated from parent ion delta masses.

Like other scoring methods, LFDR incorporates multiple scores when they are reported by a search engine. Instead of PeptideProphet's LDA or Percolator's SVM classifier, LFDR uses log-likelihood ratios generated by Naïve Bayes classifiers to discriminate between target and decoy hits. Naïve Bayes was chosen specifically for robustness to over-fitting, a frequently occurring problem when training a classifier on a subset of testing data.

Training data is selected by iteratively testing three sets of ten classifiers to hone in on the optimal number of spectra to avoid training with incorrect identifications assigned to target proteins. Then the posterior peptide probabilities are derived using LFDR estimates in a Bayesian framework. Instead of considering LFDR bins of discrete score distances, Scaffold uses variable width bins keeping the number of values in each bin constant. This gives more refined assessments of probability in score areas with more values, while simultaneously ensuring that LFDR estimates stay reasonable with fewer. Finally, a Bayesian algorithm is used to confirm peptide probabilities based on likelihoods calculated using parent mass accuracy. See [LFDR_users_meeting_2013.ppt](#)

PeptideProphet

When using PeptideProphet Scaffold determines the distributions of the scores assigned by a search engine like SEQUEST, Mascot, MaxQuant or others, which depend on the database size used for the search and the specific characteristics of the analyzed sample, see [Keller \(2002\)](#). From these distributions, Scaffold translates the search engine scores into the probabilities that a given identification is correct. Scaffold's probabilities can then be used as threshold filters, allowing the identifications to be viewed at various confidence levels.

Scaffold's method contrasts with SEQUEST's, which uses an XCorr cut-off that depends on neither database size nor sample characteristics, frequently requiring ad hoc corrections for these parameters. Scaffold's statistical approach yields more reliable estimates of the probability of a correct identification.

Scaffold's method also supplements Mascot's. Mascot provides a probability estimate based on database size, but not on sample characteristics. By incorporating the sample-specific distribution, Scaffold provides better estimates of the probability of a correct identification.

ProteinProphet

ProteinProphet groups the peptides by their corresponding protein(s) to compute probabilities that those proteins were present in the original sample, see [Nesvizhskii \(2003\)](#).

In Scaffold 4 modified weights for protein probability calculations are used in the ProteinProphet algorithm to more accurately model peptide assignments. The Similarity View has been modified to reflect these changes by reporting the peptides weights used as percentages when the User selects to group the data using the clustering algorithm.

Comparisons to increase confidence in protein identification

To increase the confidence in protein identifications, Scaffold offers a number of instructive comparisons:

- Run Scaffold's bundled version of X!Tandem, see [Validation with X!Tandem](#), on data from another search engine. Peptides found to match with both SEQUEST and X!Tandem, or both Mascot and X!Tandem, are more likely to be valid than those peptides for which the two search engines disagree.
- Compare replicates of a biological sample to see if the same proteins are identified in each. The [Samples View](#) facilitates this with a direct, side-by-side view of all samples.
- Compare replicate proteins from the same spots on different gels to see if the same proteins are identified in each of them. The [Proteins View](#) allows to sort spots according to the proteins they contain, the gel they came from, or other labels the User chooses.

Chapter 2

Identifying Proteins with Scaffold

- Compare the peptide patterns seen in each replicate. The sequence coverage shown in the [Proteins View](#) enables the User to determine at a glance whether the same peptides appear repeatedly in various samples.
- For each protein, compare the number of peptides identified. For each peptide, compare its scores from various search engines. The [Proteins View](#) lists the peptides and associated statistics.
- For peptide identifications of interest, examine the spectrum available in the [Proteins View](#). Long ladders that match ion peaks with the amino acids in the peptide sequence strongly indicate valid results, whereas large numbers of unidentified peaks do not.
- For increased confidence in Scaffold's statistical analysis examine the [Statistics View](#) to insure that statistical assumptions are met.

For information on combining multiple searches, see [Searle \(2008\)](#) and proteome-software.wikispaces.com/file/view/improving-sensitivity-by-combining-MS-MS-results.

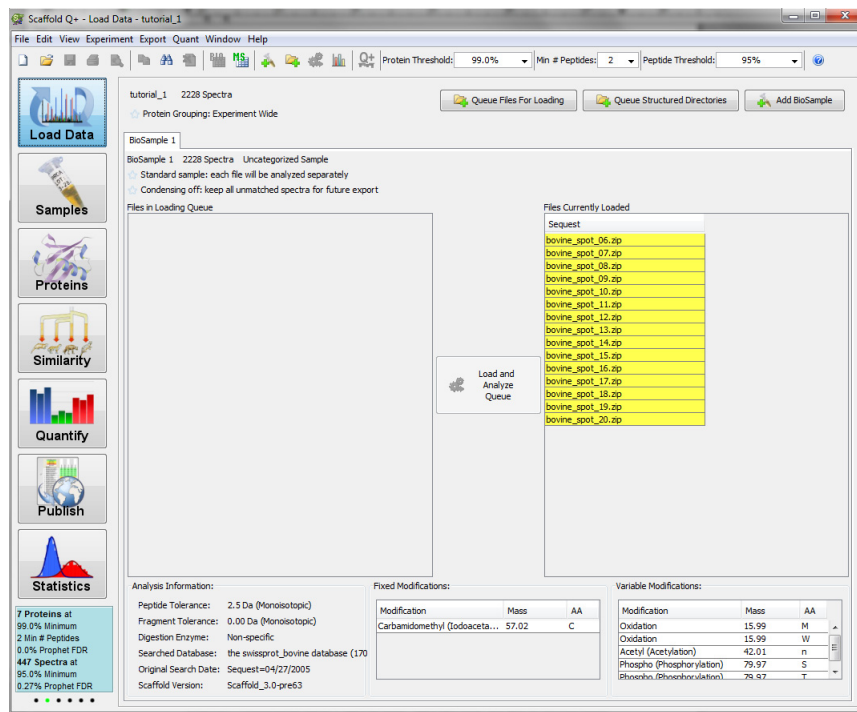
Scaffold Views

Scaffold offers both a high-level overview of the imported search results and a detailed look at supporting data, facilitating both top-down and bottom-up analysis. Scaffold presents the more detailed levels in a coherent structure, helping the User in verifying critical findings.

Load Data View

This view allows the User to load additional data, review the list of files loaded in each BioSample, edit BioSample's information, or delete already loaded MS Samples.

Figure 2-3: Load Data View



Samples View

Scaffold's [Running H/F 2](#) provides overviews that help the User make direct comparisons among categories of samples, BioSamples and MS samples. It lists and summarizes the proteins identified in each MS sample.

The list of proteins is shown summarized in two levels of hierarchy:

- **Protein groups** - groups of proteins that are associated with an identical set of peptides. They are shown collapsed and, by default, represented by the protein that has the highest

Chapter 2

Identifying Proteins with Scaffold

probability and the largest associated number of spectra. The proteins included in the group are shown in the Protein Information Pane.

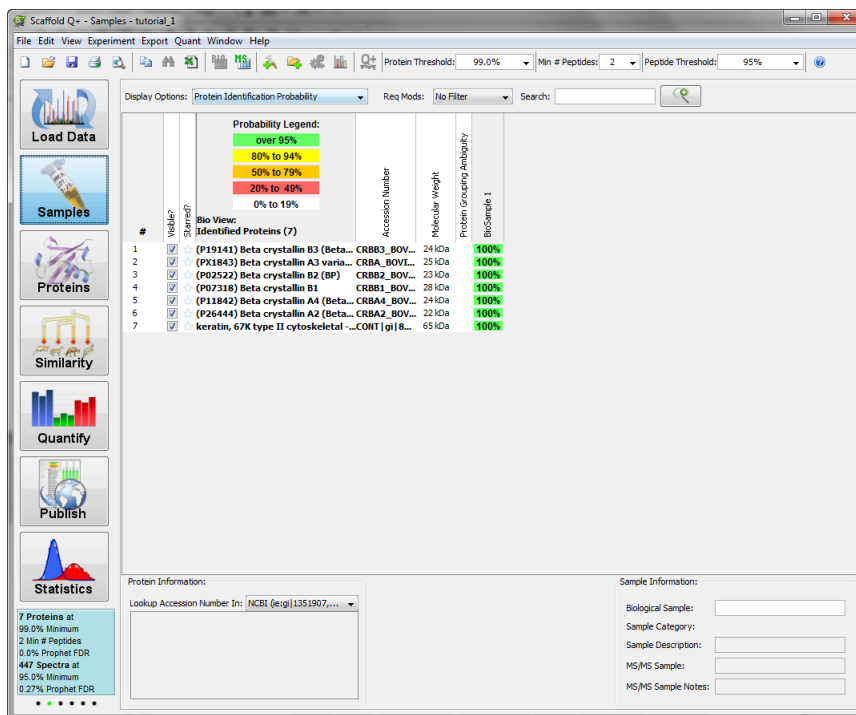
- **Protein Clusters** - sets of proteins or proteins groups created using a hierarchical clustering algorithm similar, but more stringent, to the Mascot's family clustering algorithm. Proteins or protein groups members of the cluster share some peptides but not all of them. They are by default represented by the protein that shows the highest associated probability. Clusters can be collapsed or expanded directly in the protein list.

For each protein or protein group that Scaffold identifies various Display Options are available providing different statistical information and counting options, see [“Display Options” on page 135](#).

For the highest-level overview, MS samples are grouped into BioSamples and results can be viewed collapsed in a single column summary for the entire group of MS samples, for further information go to [“MS Sample vs BioSample summarization levels” on page 124](#).

To better focus on the most useful results, confidence filters allow setting minimum standards for identification probability, or for the other available Display Options. It is possible in this way to screen out less significant findings for a shorter, higher confidence list. Or by relaxing the filters it is possible to find less-confident identifications that might be more promising for further investigation, for more information go to [“Sorting feature” on page 127](#).

Figure 2-4: Samples View

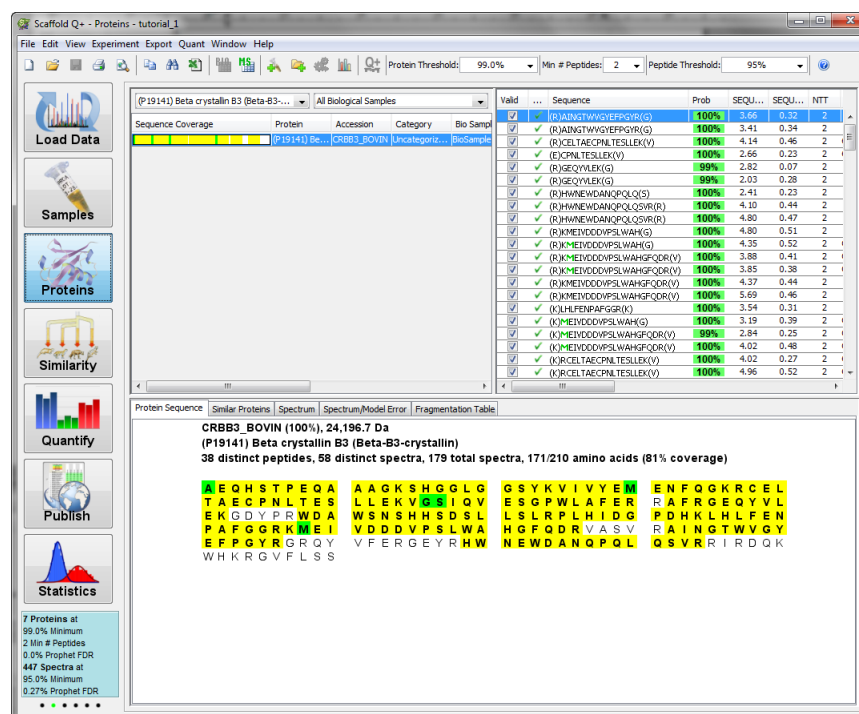


Proteins View

Scaffold's Proteins view structures a large amount of detailed information about a protein:

- Sequence coverage for this and similar proteins;
- The peptide sequence, with identified peptides highlighted in yellow and modifications highlighted in green;
- The spectra used to identify each peptide, with associated error measurements;
- The fragmentation table listing the ion fragments along with their associated peaks.

Figure 2-5: Proteins View



For each peptide, the user can also see:

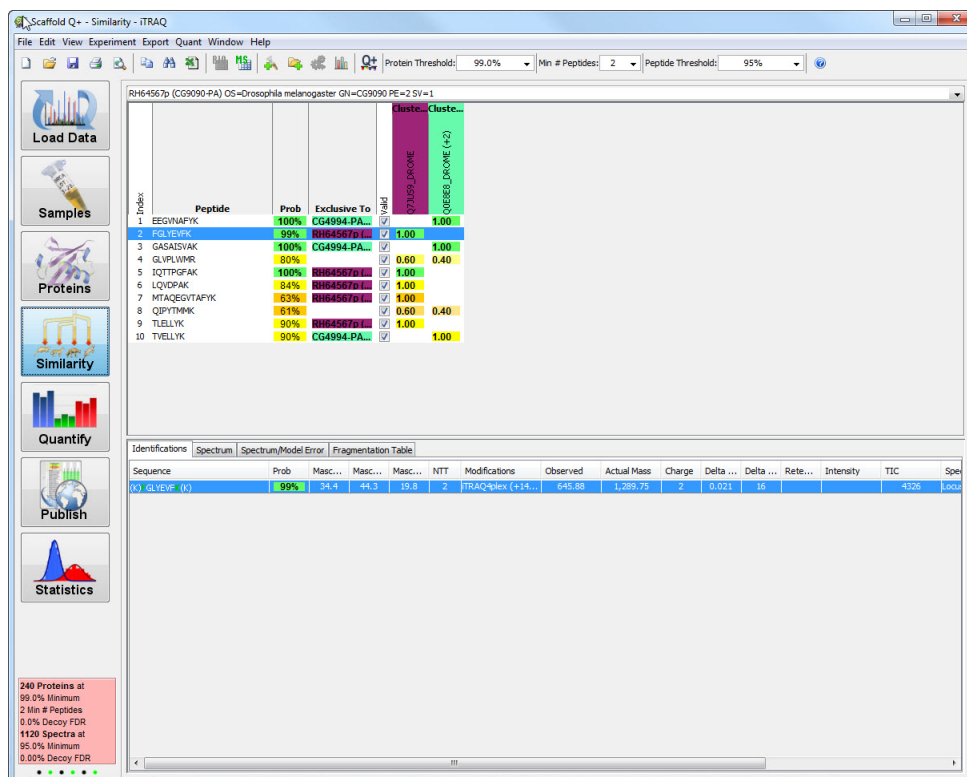
- its charge, mass, and position in the peptide sequence;
- associated confidence scores from other search engines;
- modifications, if any.

With all this information available at a glance or at a mouse click, the user can have confidence in his/her results and organized evidence to document the various findings.

Similarity View

The Similarity View allows to see how the different peptides are shared in the various hierarchal grouping levels.

Figure 2-6: Similarity View



For each peptide, the corresponding proteins to which it could belong are listed on the right.

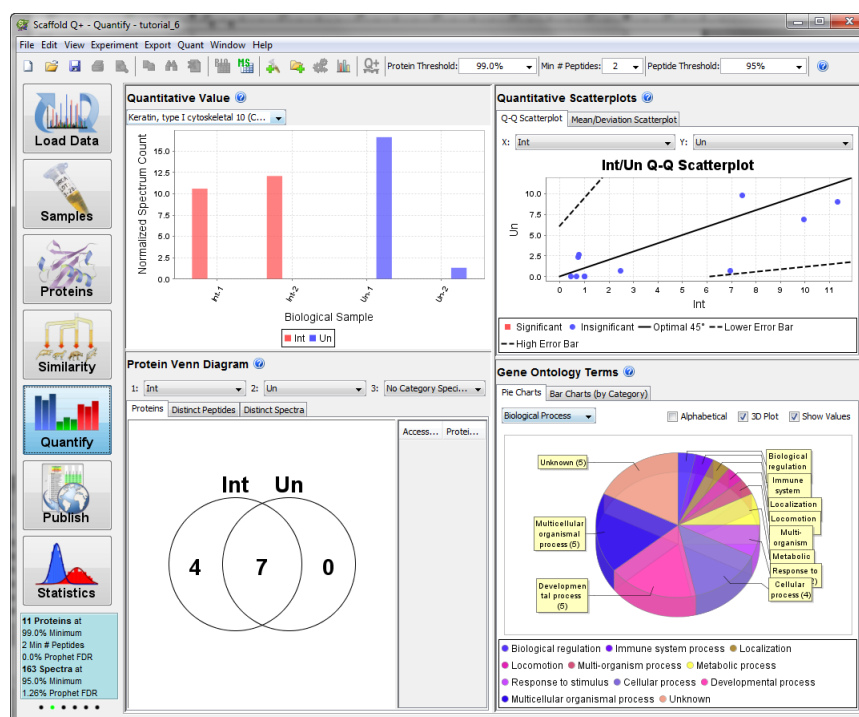
- The user can “check” or “uncheck” the valid box for a peptide sequence. Unchecking the box removes that peptide from Scaffold's probability calculations.
- Peptides identified in particular protein groups are color coded to match their protein group

Quantify View

In the Scaffold's [The Quantify View](#) the User can look at spectral count numbers for BioSamples along with their associated errors and compare spectral counts between different BioSamples and categories. The information is organized in the following panes:

- **The Quantitative Value pane** - where spectrum counts can be viewed for a selected protein. A drop down list allows the User to choose which protein's spectrum counts are displayed.
- **The Quantitative Scatterplots pane** - where the User can check the degree of error associated with the spectral count measurements.
- **The Venn diagram pane** - where the User can see the relationship between or among categories of proteins, exclusive distinct peptides, or exclusive distinct spectra identifications.
- **The Gene Ontology Terms pane** - where the User can see a pie chart displaying the GO terms for the overall Scaffold experiment or select the Bar Charts tab to view the GO terms by category.

Figure 2-7: Quantify View



Publish View

The Publish View provides detailed information about the experiment in general. The view includes two tabs:

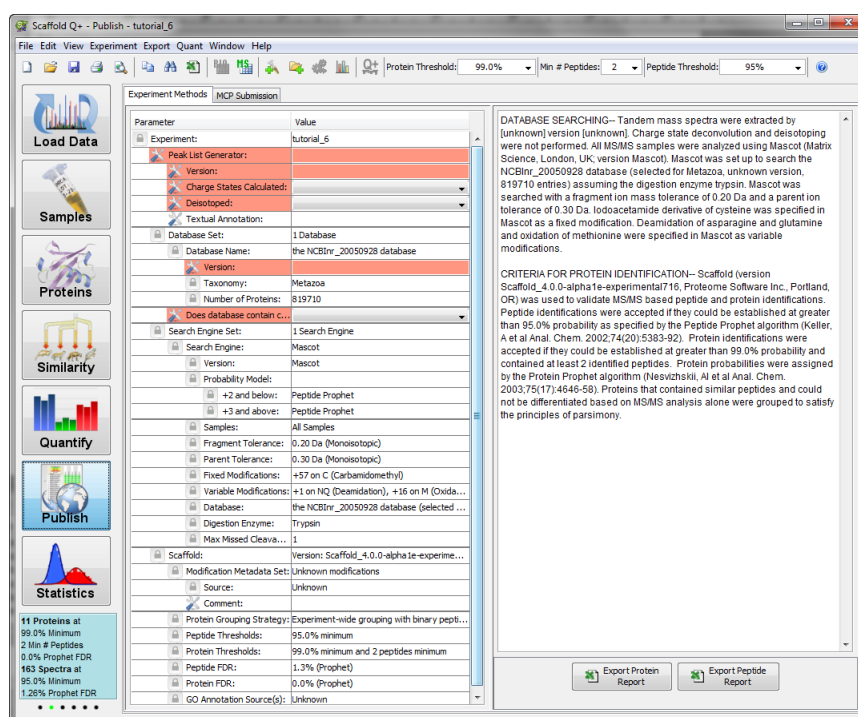
- **The Experimental Methods tab** - which describes the parameters used when performing the experiment. This is information about the experiment typically required by the major proteomics journals like:

Chapter 2

Identifying Proteins with Scaffold

- Molecular & Cellular Proteomics,
- Proteomics
- Journal of Proteomics Research
- **The MCP Submission tab** - which is a checklist that will lead you through the process of submitting all the supplemental material needed for publication in these proteomics journals. This procedure packages the data and stores it on the file sharing service Tranche.

Figure 2-8: Publish View



Using the *mzIdentML* exports, it is now possible to submit data analyzed through Scaffold to the PRIDE public data repository.

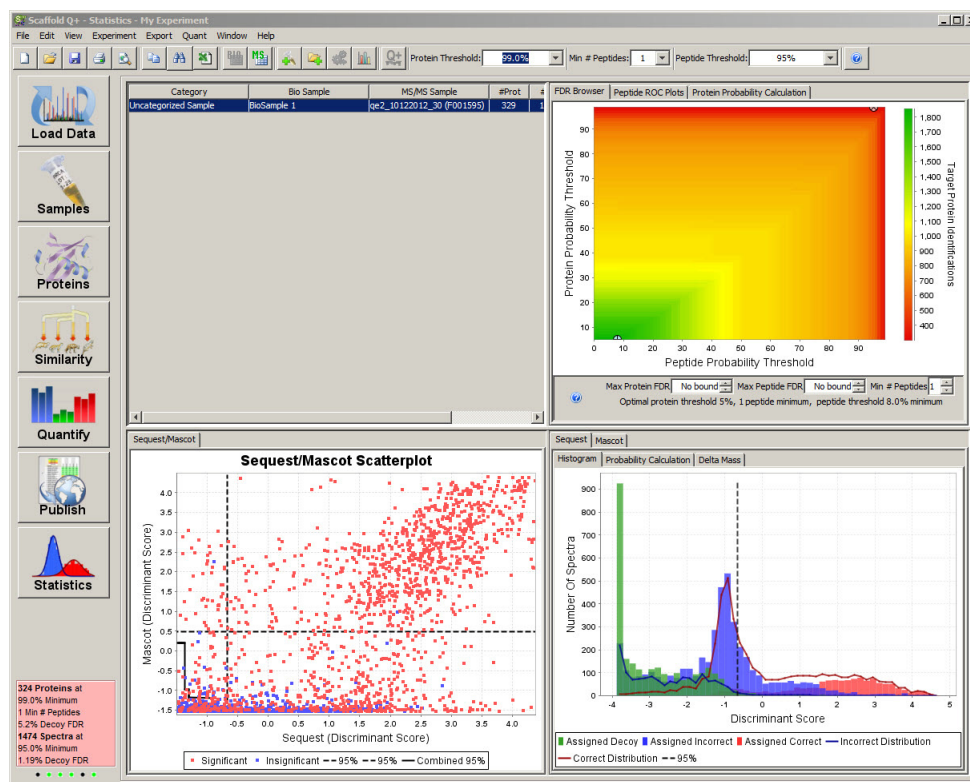
Statistics View

The Statistics view displays:

- statistical information for each MS sample in the analysis;
- the relationship between peptide and protein probabilities;
- a histogram demonstrating correct and incorrect peptide assignments;

- a scatterplot comparing two or more search engine results (this will only displayed when multiple search engines were used).

Figure 2-9: Statistics View



The Statistics view provides a way to verify that the underlying assumptions of Scaffold are met. If data meets Scaffold's assumptions the user can have confidence in the analysis results.

Chapter 2

Identifying Proteins with Scaffold

Chapter 3

Loading Data in Scaffold

Scaffold can import and analyze data produced by a variety of search engines. All results can be freely mixed in a given experiment or a given BioSample as long as the different data files have been searched against the same database.



When multiple search engine results are included in the same BioSample, Scaffold recognizes this and groups the different outputs together as one MS sample.

This chapter covers the following topics:.

- [“The Loading Wizard” on page 38](#), helps a new User go through an example that shows how to load files in Scaffold and describes the different steps within the loading Wizard
- [“Modify make up of BioSamples” on page 50](#), which shows how to adjust the description and name of the loaded samples and delete some of the files already loaded
- [“Specifying the FASTA database” on page 53](#), which shows how to load and parse databases in Scaffold through an example
- [“Validation with X!Tandem” on page 57](#), which explains how to run X!Tandem through Scaffold

The Loading Wizard

To familiarize the new User with the Scaffold Wizard we have developed a short exercise that shows step by step how to load a number of example files into Scaffold using the loading Wizard. Details about the files used in this exercise are provide in the boxes below.

Each page in the Wizard points the User to a different task which is detailed in the following procedure:

- [Select Quantitative Technique](#)
- [New BioSample](#)
- [Queue files for loading](#)
- [Queue more files for loading](#)
- [Add another BioSample](#)
- [Load and Analyze Data](#)



The following procedure is written from the perspective of loading either:

- the SEQUEST data in folder **tutorial_3seq** and related FASTA file (swissprot_bovine.fasta) that are available at: http://www.proteomesoftware.com/products/data/sequest_tutorial.zip.

Or from the perspective of loading:

- the Mascot data in folder **tutorial_3mas** and related FASTA file (control_sprot.fasta) that are available at: http://www.proteomesoftware.com/products/data/mascot_tutorial.zip.

To carry out this procedure using this sample data, you have to first extract the contents of the zip file.



If you are following this procedure using the SEQUEST data files, to shorten the time Scaffold takes to access SEQUEST's numerous subfolders at the operating system level (outside of the Scaffold program), navigate to the folder in which you placed tutorial_3seq and open it, briefly viewing the sub-folders within it.

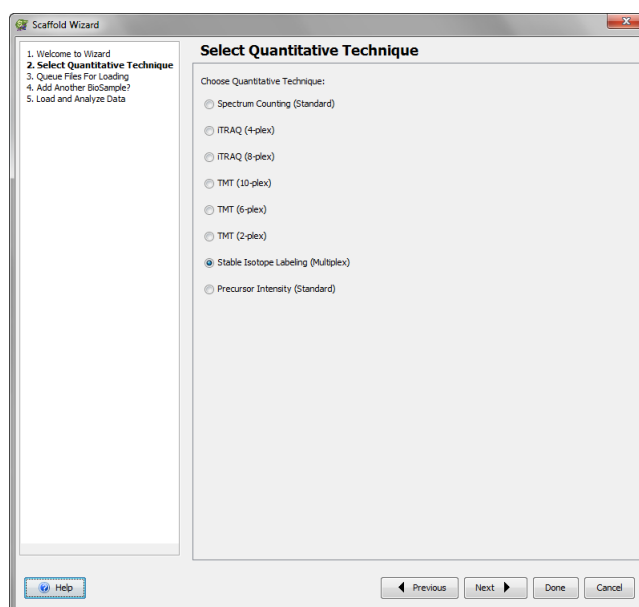


*When you see Mascot output over the web, you're viewing HTML summaries of results. These are not valid input for Scaffold. Scaffold requires *.DAT files from Mascot. The path and filename is usually visible as the last part of the URL in the address field of the browser displaying the results page, after the file=*

Select Quantitative Technique

1. Open Scaffold.
2. In the Welcome to Scaffold or Scaffold Q+ or Scaffold Q+S window, click **New**.
The Scaffold Wizard, Welcome to Wizard page opens, click **Next** to go to the **Select Quantitative Technique** page if you are running Scaffold Q+ or Q+S.
3. Specify how Scaffold Q+ or Scaffold Q+S is to quantitatively treat the samples.

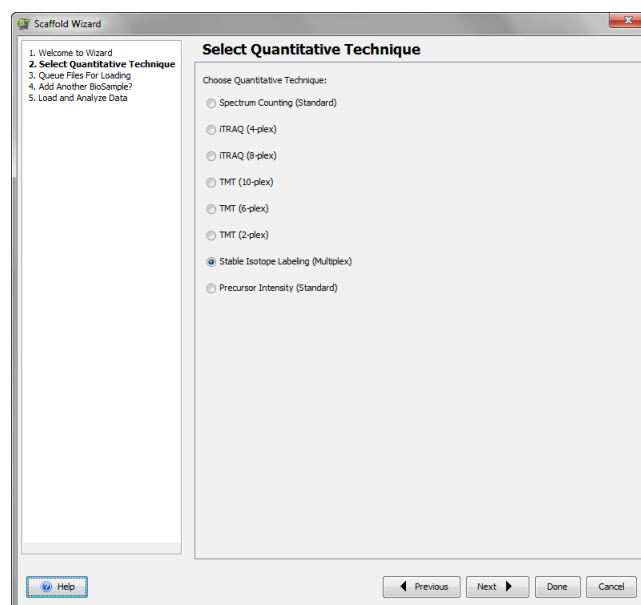
Figure 3-1: Scaffold Wizard, New Quantitative Technique page in Scaffold Q+



Chapter 3

Loading Data in Scaffold

Figure 3-2: Scaffold Wizard, New Quantitative Technique page in Scaffold Q+S

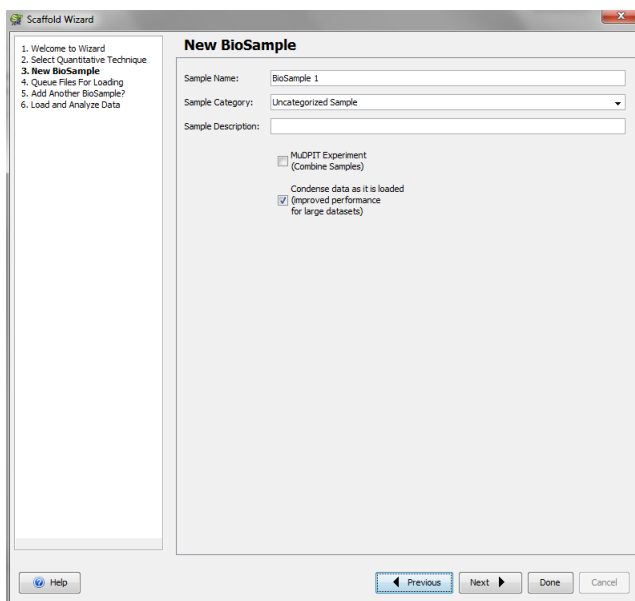


If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, then select Spectrum Counts.

4. Click **Next**.

The Scaffold Wizard, New BioSample page opens.

Figure 3-3: Scaffold Wizard, New BioSample page



5. Continue to “New BioSample” below.

New BioSample

1. Enter a sample name, and optionally, enter a description that further clarifies or explains the sample.



If you are carrying out this procedure using the sample tutorial_3seq data provided by Proteome Software, name the new BioSample bovine 1ens and the new category 1ens. Because these names appear as column headings in the Samples View, it's helpful to choose short ones. When there's more to remember, enter it in the Sample Description field.



If you are carrying out this procedure using the sample tutorial_3mas data provided by Proteome Software, name the new BioSample c1 and the new category control.

2. Do one or both of the following as needed:
 - To discard 0% probability spectra and decrease the time required to load the data, select Condense data as it is loaded.
 - If the loaded data is MuDPIT data, then select MuDPIT Experiment. Scaffold combines all the MS samples for the sample.

Chapter 3

Loading Data in Scaffold

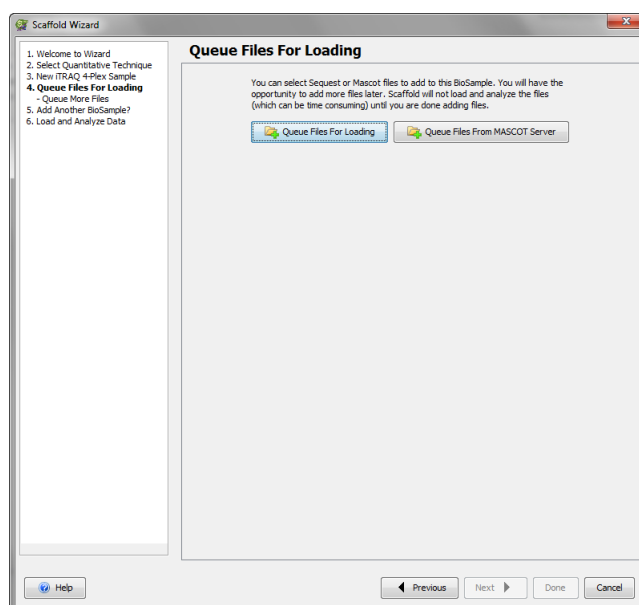


If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, leave both these boxes unchecked.

3. Click **Next**.

The Scaffold Wizard, Queue Files for Loading page opens.

Figure 3-4: Scaffold Wizard, Queue Files for Loading page



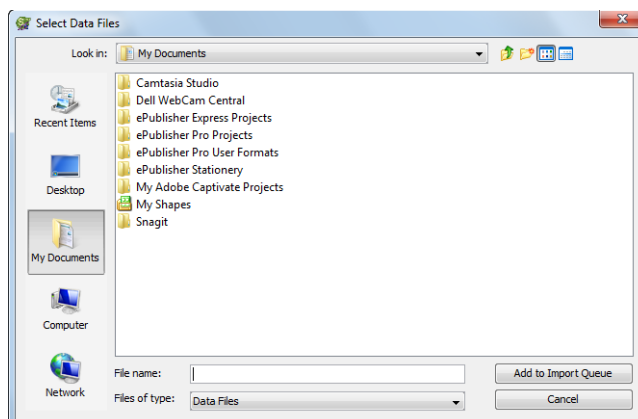
4. Continue to [“Queue files for loading” on page 42](#).

Queue files for loading

1. Click Queue Files for Loading.

The Select Data Files dialog box opens.

Figure 3-5: Select Data Files dialog box



2. Navigate to the directory in which you saved your sample data set and FASTA database, select the sample data set, and then click Add to Import Queue.



If you are carrying out this procedure using the sample tutorial_3seq data provided by Proteome Software:

- Open the folder tutorial_3seq.
- Select several or all of all the sub-folders bovine_spot_01 through bovine_spot_20.

Each of these folders represents one mass spectrometry sample holding data from the corresponding spot in the 2D gel.



*Note: if you open one of those folders and just select one of the *.OUT files it contains, Scaffold will automatically load all of the files in that folder. That is because SEQUEST places the information related to one MS sample search in numerous separate files within the folder.*



If you are carrying out this procedure using the sample tutorial_3mas data provided by Proteome Software:

- Open the folder tutorial_3mas.
- Select the file control_01.dat

The Select Data file dialog box closes, and you return to the Scaffold Wizard, Queue Files for Loading, Queue More Files.

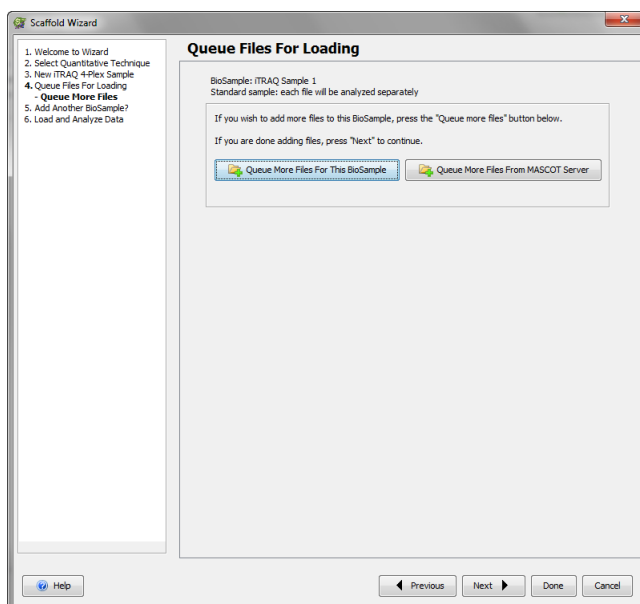
3. Queue More Files

The page prompts you to load additional data files for the current BioSample.

Chapter 3

Loading Data in Scaffold

Figure 3-6: Scaffold Wizard, Queue More Files for Loading page



4. Continue to [“Queue more files for loading”](#) on page 44.

Queue more files for loading

1. If you have more data files to load for the current *BioSample*, then do the following for each set of these data files; otherwise, continue to [Step 2](#).



If you are carrying out this procedure using the sample tutorial _3seq data provided by Proteome Software, do not add more file to the BioSample.



If you are carrying out this procedure using the sample tutorial _3mas data provided by Proteome Software, each file is loaded in its own BioSample so continue to [Step 2](#)

- Click Queue More Files For This BioSample.
- Repeat [“Queue files for loading”](#)

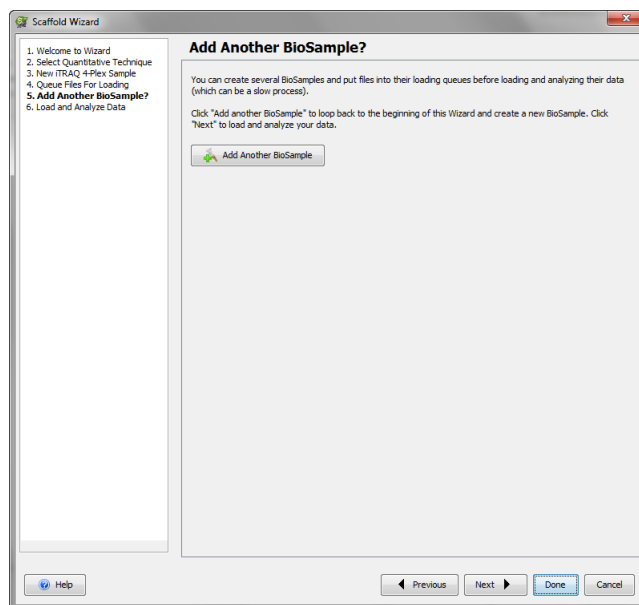


NOTE: Selecting category descriptions whenever possible from the drop-down list of other names you have used will make sure you don't incorporate unintentional small differences in naming, which would prevent proper sorting in the Samples view.

2. Click Next.

The Scaffold Wizard, Add Another BioSample? page opens.

Figure 3-7: Scaffold Wizard, Add Another BioSample? page



3. Continue to [“Add another BioSample”](#) on page 46.

Add another BioSample

1. Do one of the following:
 - If you have other *BioSamples* that are to be analyzed, then for each of these BioSamples, click Add Another BioSample to return to page 2 of the Scaffold Wizard, cycle through the wizard to add the sample, and then click Next.
 - If you do not have other BioSamples that are to be analyzed, then click Next.



If you are carrying out this procedure using the sample tutorial_3seq data provided by Proteome Software, then click Next.



If you are carrying out this procedure using the sample tutorial_3mas data provided by Proteome Software, Repeat the process for the second replicate: naming the new BioSample c2 and choosing the same category description of control from the drop-down list.

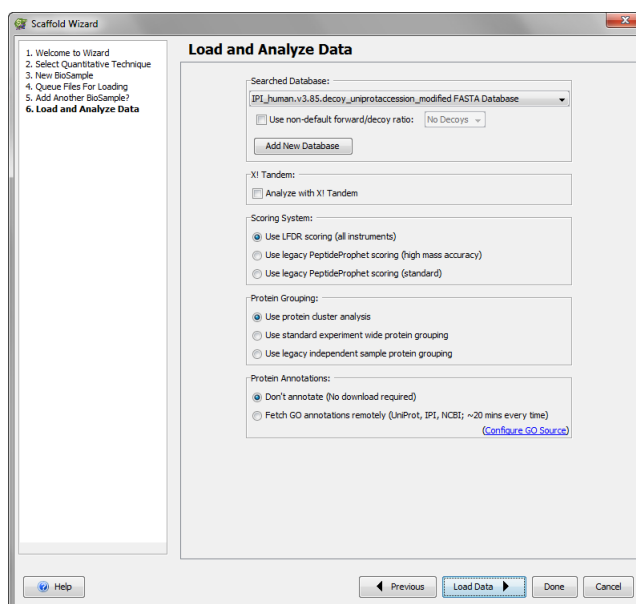
Repeat the procedure starting from “[Add another BioSample](#)” on page 46 until you have added all the samples you wish. Then Click Next to go to “[Load and Analyze Data](#)” on page 46

Load and Analyze Data

The Scaffold Wizard, Load and Analyze Data page opens. The page is divided into various panes. Each pane contains options for customizing the way Scaffold analyzes the data during the loading phase:

- [Searched Database Pane](#)
- [Analyze with X!Tandem Pane](#)
- [Scoring System Pane](#)
- [Protein Grouping Pane](#)
- [Protein Annotations pane](#)

Figure 3-8: Scaffold Wizard, Load and Analyze Data page



- Continue to [“Specifying analysis options and analyzing the data” on page 47.](#)

Specifying analysis options and analyzing the data

Searched Database Pane

This pane allows the User to select or import the database used to create the data files loaded in Scaffold. Databases previously loaded will appear listed in a pull down list. To add new databases the User can click the Add New Database button. For more detailed information continue to [“Specifying the FASTA database” on page 53.](#)

Analyze with X!Tandem Pane

Selecting this option runs an additional database search, an X!Tandem search, on the loaded data with variable modifications chosen by the User. This operation improves protein identifications but significantly increases analysis times.

For more information continue to [“Validation with X!Tandem” on page 57](#)



If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, then select run with X!Tandem.

Scoring System Pane

This pane offers different post processing scoring algorithms to apply when Scaffold analyzes the imported data:

- **Use LFDR Scoring** - Algorithm for assessing the confidence level of the identified peptides. Based on a Bayesian approach to LFDR (Local False Discovery Rate), this algorithm, introduced with Scaffold version 4, is particularly effective for QExactive and high mass accuracy data, see [LFDR-based scoring system](#).
- **Use Legacy PeptideProphet Scoring (high Mass Accuracy)** - This option will use the standard PeptideProphet algorithm developed in Scaffold version 3 and older together with the high mass accuracy option
- **Use Legacy PeptideProphet Scoring (Standard)** - Standard PeptideProphet with no high mass accuracy.

For references and information about the scoring algorithms used in Scaffold see [Algorithms References](#).



When the data set to be analyzed was not searched using the decoy option or against a decoy concatenated database, the Legacy PeptideProphet Scoring high Mass accuracy option will be automatically selected.

Protein Grouping Pane

This pane shows the available grouping analysis options performed by Scaffold over the list of identified proteins.

- **Use protein cluster analysis**- Since Scaffold 4, a new hierarchical grouping level was added above the Scaffold standard protein grouping. While similar to the Mascot's hierarchical family clustering, Scaffold 4 clusters are created using added stringencies that often succeed in separating proteins into sets of biologically meaningful isoforms. Each cluster showed in the Samples View can be expanded or collapsed. The clusters sub-menu contains options for expanding or collapsing all of the protein clusters displayed in the Samples View.



The initial default option is Use Protein Cluster Analysis

- Use standard experiment wide protein grouping - When selected Scaffold groups proteins across all MS samples and BioSamples.
- **Use legacy independent sample protein grouping** - When selected Scaffold groups proteins only within each MS sample. Each MS samples appears as if it was loaded independently.



If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, select the option "Use standard experiment wide protein grouping".

For more information on the grouping and clustering algorithms used in Scaffold see [Chapter 8, “Protein Grouping and Clustering,”](#) on page 151.

Protein Annotations pane

Included options for searching the Gene Ontology annotations, GO terms, during loading:

- **Don't Annotate**
- ***Fetch Go annotations remotely.*** If the GOA database is not configured, the option will appear grayed out. For activation click the link *Configure GO Source* and select a GOA database from the [Edit GO Term Options](#), GO Annotation Databases pane. If the database you are searching is not available click New database and import the GOA database of your interest.



If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, then select Don't Annotate.

2. If you have selected to run X!Tandem continue to [“Validation with X!Tandem”](#) on page 57
3. Once all the options have been properly checked, click:
 - Load and Analyze Data if X!Tandem was not selected.

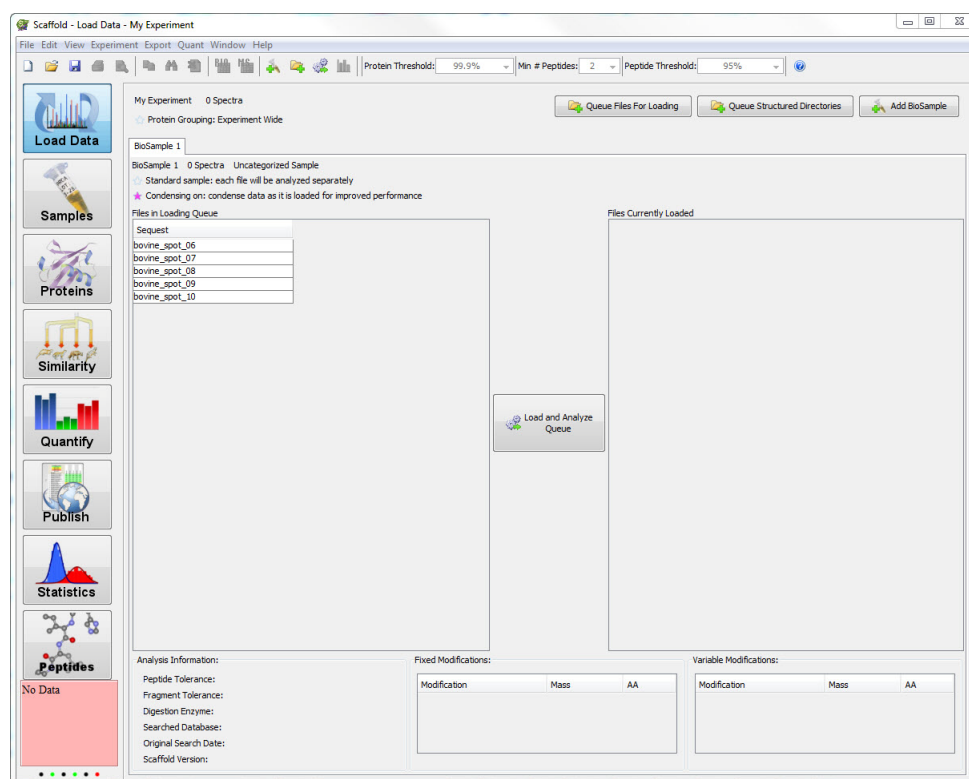
A message opens, indicating that the data is being loaded and analyzed. After the analysis is complete, the data opens in the Samples View.

Continue to [“Modify make up of BioSamples”](#) on page 50.

Modify make up of BioSamples

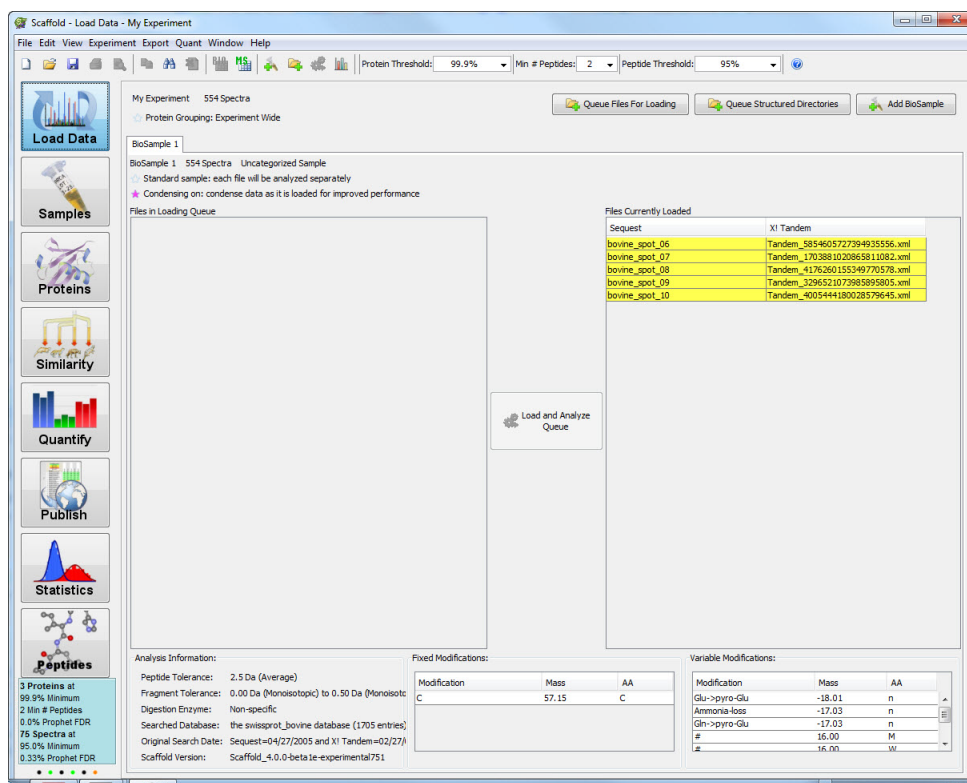
When the files have been loaded, they move from the import queue on the left side of the Load data View table to the ready pane on the right.

Figure 3-9: Load Data View Files in Import Queue



When they've been analyzed, Scaffold highlights them in yellow and then switches to the Samples view

Figure 3-10: Load Data View Files in Import Queue



At any time the User can remove files from either the Loading Queue or the Ready Pane.

If the User should load a file by mistake, or wish to change the make-up of a BioSample he/she needs to do the following:

1. Click the Load Data icon to go to the Load Data View
2. You will see the files loaded and analyzed on the right. Click on one to select it.



Note that you select the entire MS sample, including both the SEQUEST (or Mascot) and X! Tandem runs.

3. Click the right mouse button. You will see a single menu item —Remove Selected Samples. Click it.
4. A second dialog asks you to confirm the removal. For now, click Don't Remove

Save the Experiment

To Save the experiment:

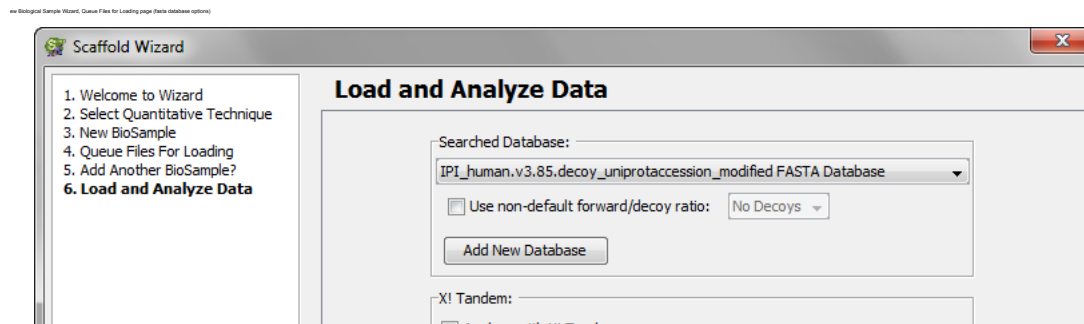
Chapter 3

Loading Data in Scaffold

1. Go to the File menu and select Save
2. Navigate to the folder in which you wish to save these tutorials, and enter the filename tutorial_3seq. Scaffold appends the suffix .SF3 to your experiment files.

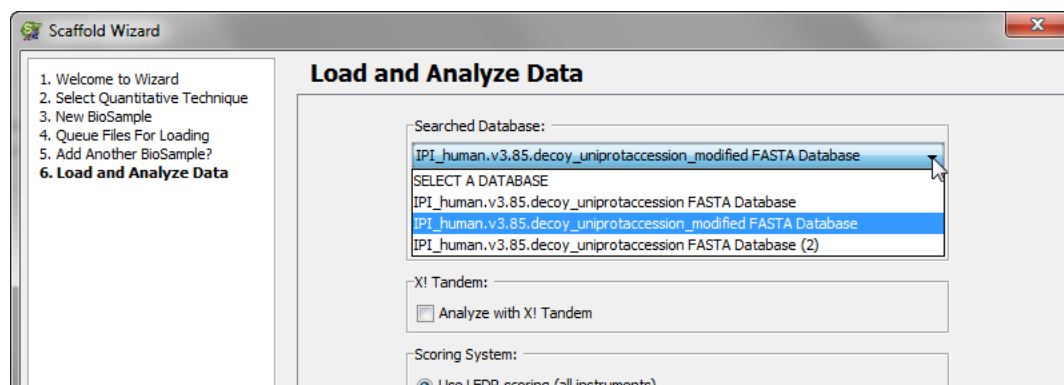
Specifying the FASTA database

Figure 3-11: Searched Database Pane



1. Specify the FASTA database that is associated with these sample files. You can select from a list of existing FASTA databases shown in the pull down menu, or you can add a new FASTA database.
 - Just select a database from the existing list-+.

Figure 3-12:



- If you are adding a new database, continue to [Step 2](#).



If you are carrying out this procedure using the sample tutorial [3seq](#) data or the sample tutorial [3mas](#) data provided by Proteome Software, then continue to [Step 2](#).



If the FASTA database selected is not identical to the external protein database, including the version, that you used for searching your experimental data, then the protein sequence and molecular weight might not be available later in the Protein View

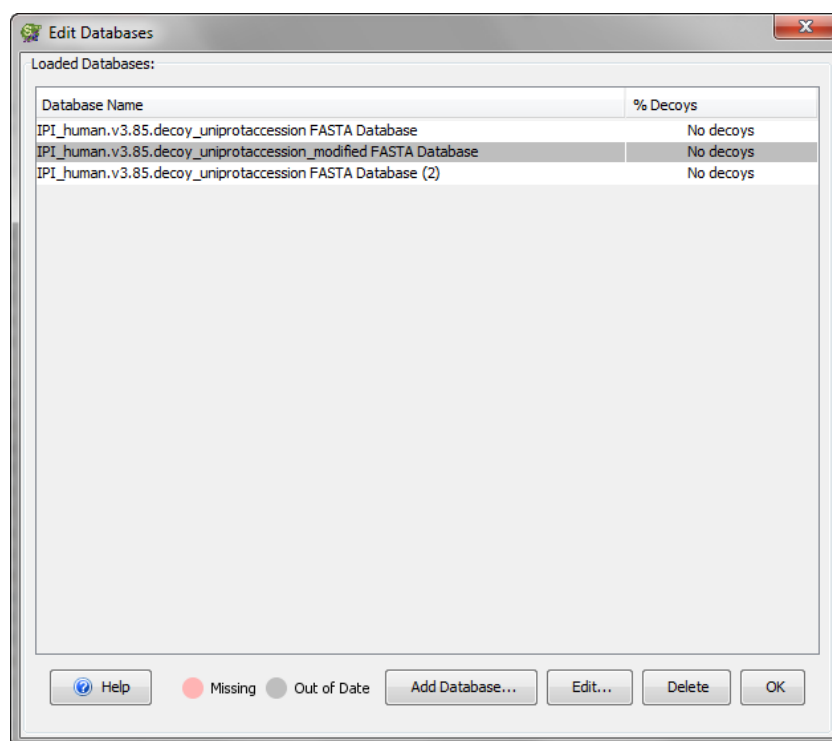
2. Click Add New Database.

The Edit Databases dialog box opens.

Chapter 3

Loading Data in Scaffold

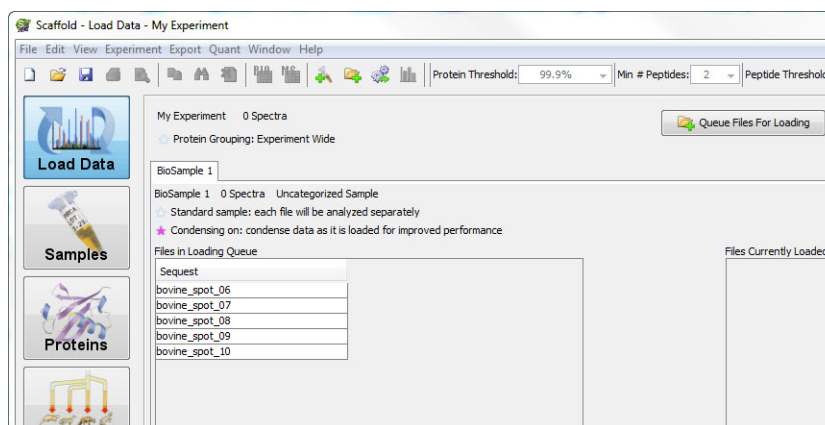
Figure 3-13: Edit Databases dialog box



3. On the Edit Databases dialog box, click New Database.

The Open FASTA Database dialog box opens.

Figure 3-14: Open FASTA Database dialog box



4. Navigate to the directory in which you saved your sample data set and FASTA database, select the FASTA database, and then click Open.



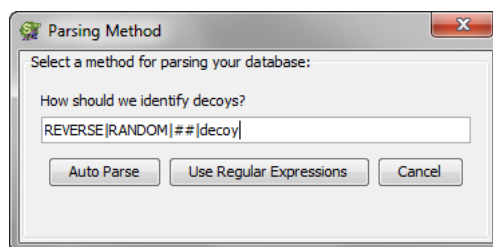
If you are carrying out this procedure using the sample tutorial_3seq data provided by Proteome Software, navigate to where you have saved the subset FASTA data base `swissprot_bovine.fasta`.



If you are carrying out this procedure using the sample tutorial_3mas data provided by Proteome Software, navigate to where you have saved the subset FASTA data base `control_sprot.fasta`.

The Parsing Method dialog box opens. You use the options on this dialog box to select the parsing rules that display protein accession numbers and protein descriptions in the correct format.

Figure 3-15: Parsing Method dialog box



5. Do one of the following:

- Click Auto Parse to have Scaffold decide on the parsing rules to use. If you are parsing a database that contains decoys make sure the decoy identification tag is included in the How should we identify Decoys? list shown in the dialog box.



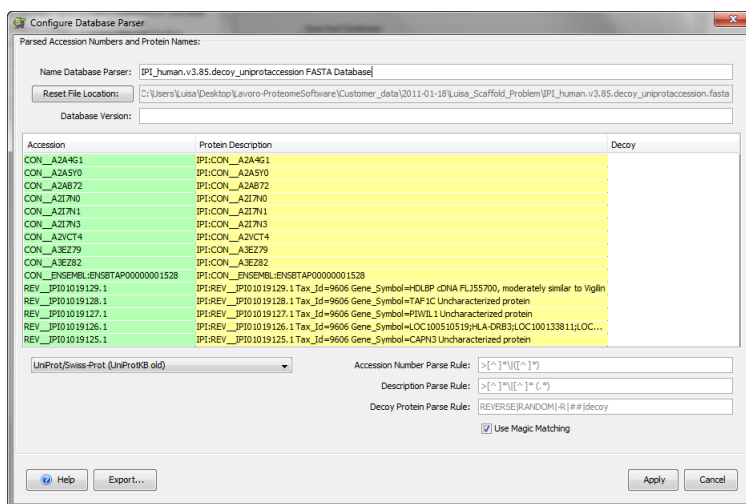
Auto Parse is the preferred method for parsing the database. If you are carrying out this procedure using the sample tutorial_3seq data or the sample tutorial_3mas data provided by Proteome Software, then select Auto Parse.

- Click Use Regular Expressions to open the Configure Database Parser dialog box and select a specific pre-configured parsing rule to use, or create your own parsing rule. See [Figure 3-16 on page 56](#).

Chapter 3

Loading Data in Scaffold

Figure 3-16: Configure Database Parser dialog box



6. After the parsing rules are applied, you return to the Edit Databases dialog box with the correct database selected. Click OK.
7. Continue to [“Load and Analyze Data” on page 46](#).

Validation with X!Tandem

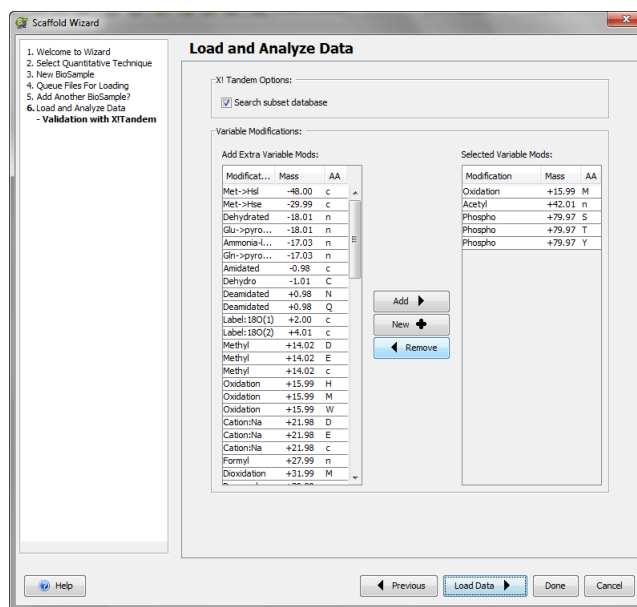
Each search engine uses its own algorithms to identify proteins. Identification confidence is higher when multiple algorithms find the same protein. Likewise, knowing which protein IDs are not confirmed by a second search engine lets the User screen out some false positives. This means that adding X!Tandem results to previous output files gives more confident protein identifications. It will, however, increase processing time.

X!Tandem runs quickly relative to SEQUEST or Mascot, but the User should expect searching Swiss-Prot with several hundred thousand spectra to take a large amount of time and searching a huge database, such as the NR database, will take even much longer.

To include X!Tandem results as part of a Scaffold experiment the User has to check the box labeled Analyze with X!Tandem in the [Load and Analyze Data](#) page of the Loading Wizard. X!Tandem runs with the same parameters as the original loaded search files but variable modifications might be added to the ones already included.

1. The Scaffold Wizard, Load and Analyze - Validation with X!Tandem page opens

Figure 3-17: Scaffold Wizard, Load and Analyze - Validation with X!Tandem



The Wizard **Validation with X!Tandem** page includes two panes:

- **X!Tandem Options pane**

Since for large databases the X!Tandem search can take a long time the option **Search subset database** was added to minimize its execution time. Checking this box means that X!Tandem searches only the subset of the proteins that were previously found with the original search engine.

For example, suppose the original SEQUEST search against a million protein NR database found 100 proteins. The subset X! Tandem search will now be against only 100

proteins instead of a million. In this case that particular one step of X! Tandem will go thousands of times faster. But the X! Tandem refinement steps which search for modifications will not go any faster. If there are a huge number of spectra, this refinement step will still take considerable time. What this all means is that the X!Tandem search will be speedier, but how much speedier depends upon the number of spectra and size of the FASTA database.

- **Variable Modifications pane**

From the input files Scaffold reads the parameters used to search a database by the search engine that produced the files. The parameters include instrument mass error tolerances, digestion enzymes, and fixed and variable modifications. Scaffold then passes these parameters to X!Tandem when it is run.

Of all the parameters passed onto the X!Tandem search the User can only modify the list of variable modifications. The variable modifications already present in the original search files readily appear listed in the **Selected Variable Mods** table located on the right side of this pane. The **Add Extra Variable Mods** table shows a list of standard UNIMOD variable modifications that can be added to the **Selected Variable Mods** table.

Between the two tables there are three functional buttons which allow to **Add** or **Remove** a variable mod from the **Selected Variable Mods** table or create a **New+** custom variable mod in the **Add Extra Variable Mods** table, see [Build A Modification](#).

Selecting more variable modifications may increase the number of peptides identified. It will surely increase the run time for X!Tandem's analysis. If many modifications are chosen, it will take many times longer to execute.

Note: When a peptide starts with E or Q, X!Tandem automatically checks for the formation of pyroglutamic acid, i.e., the loss of water or ammonia, respectively. This modification happens spontaneously in solution and failure to test for it can result in missing significant peptide hits. The analogous reaction for iodoacetimide blocked cysteine (loss of ammonia) is also considered. This modification is considered to be an N-terminal modification only, so it does not affect any potential modifications specified for Q, E or C. More information is available at <http://thegpm.org/TANDEM/api/rpmm.html>

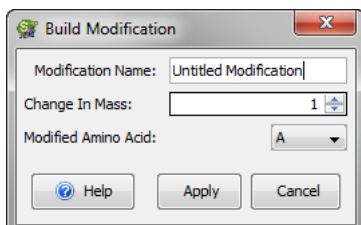
The modification tables can be sorted by clicking on the header for any column.

2. Add or remove modifications to the list already included in the original search data by using the arrows between the two lists.
3. Click Load data to start the analysis.
4. Continue to [“Modify make up of BioSamples” on page 50](#).

Build A Modification

If the User wants to search with X!Tandem using a variable modification not included in the **Add Extra Variable Mods** table he/she can define the new mod choosing **New+** on the Validation with X!Tandem Wizard page and bring up the Build Modification dialog.

Figure 3-18: Build A Modification dialog



- **Modification name** - Name that will be used in the Proteins View, Peptide pane and in the Spectrum Report for this modification. This name is saved with the Scaffold file.
- **Change in Mass** - The mass difference in amu due to this modification. Even though the modifications are only displayed as with one place after the decimal point on the modifications list, the mass is stored with the accuracy that was entered when defined.
- **Modified Amino Acid** - Pull down list with possible amino acid choices. Custom defined modifications can only apply to one amino acid. If the defined modification applies to several amino acids, the User has define several modifications accordingly.

Chapter 3

Loading Data in Scaffold

Chapter 4

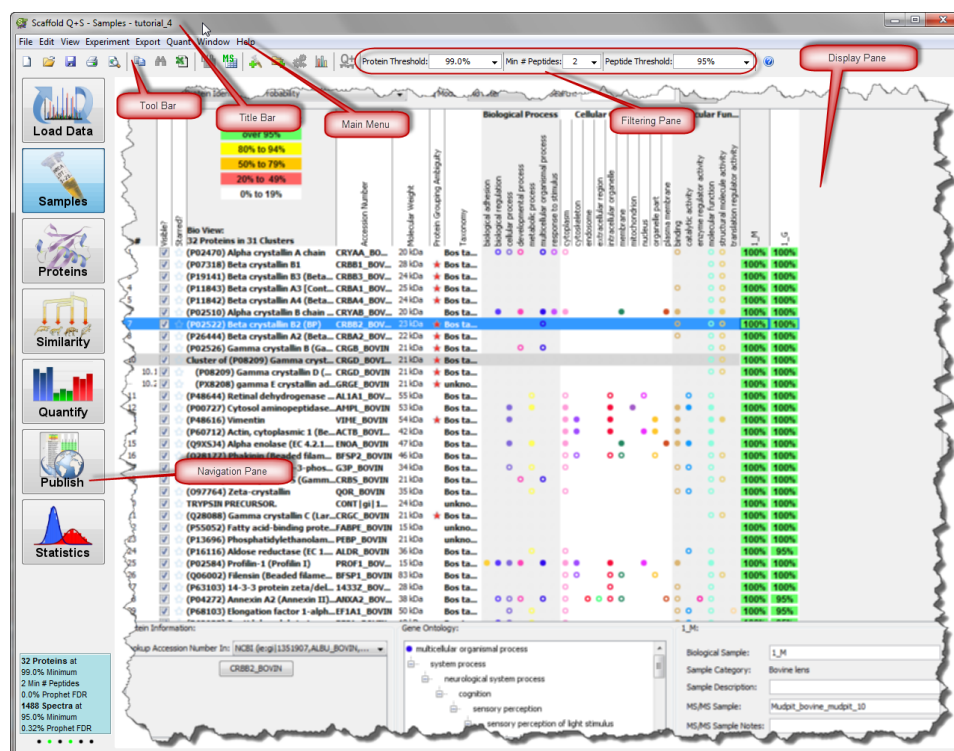
The Scaffold Window

The Scaffold application is built around a main general window containing a number of different views. Each view provides a particular perspective to look at the loaded data in the experiment. There are a number of tools available in all views and specific tools that help navigate within a selected view.

The window major components are:

- The “Title bar” on page 62
- The “Main menu commands” on page 63
- The “Tool-bar” on page 98
- The “Filtering pane” on page 100
- The “Navigation pane” on page 101
- The “Display pane” on page 104.

Figure 4-1: Scaffold window



Title bar

Figure 4-2: Title bar



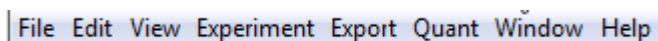
Depending on the type of license acquired either “Scaffold”, “Scaffold Q+” or “Scaffold Q+S” is always shown in the title bar at the top of the Scaffold window, together with the Scaffold icon. Additional text is displayed depending on the actions that the User is currently carrying out in Scaffold. For example, if the user has opened a file, then - <Experiment name> is also displayed in the title bar.



*The version of Scaffold in use is not displayed in the Title bar. The user must go the **Help > About** option in the main menu to determine the version number. See [“Main menu commands”](#) below.*

Main menu commands

Figure 4-3: Scaffold Main Menu

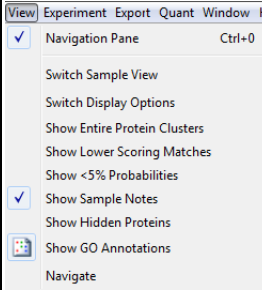
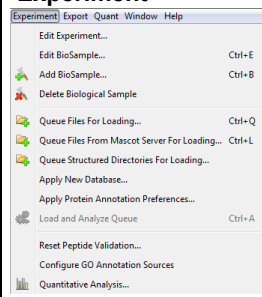


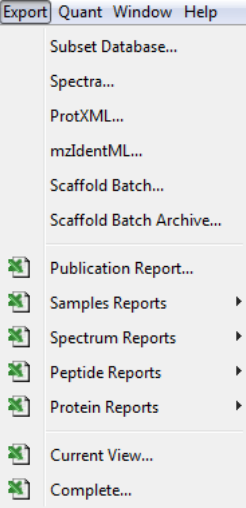
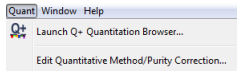
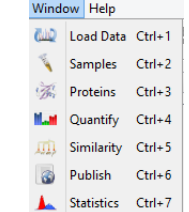
The Scaffold main menu is set up in a standard Windows menu format with menu commands grouped into menus (File, Edit, View, Experiment, Export, Quant, and Help) across the menu bar. When loading Waters IdentityE type of data an extra menu IdentityE appears after the Help menu. Some of these menu commands are available in other areas of the application as well.

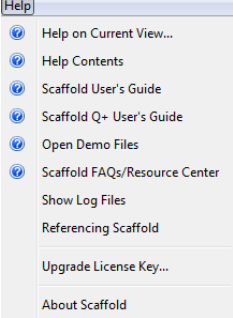
Menu	Menu Commands
File 	<ul style="list-style-type: none"> • New—Initializes a Wizard which guides the User through the loading phase of the search data files in Scaffold. See The Loading Wizard • Open—Opens a saved Scaffold experiment file, *.SF3, through a file browser. • Merge—Merges *.SF3 files in one single Scaffold experiment. See Merge. • Close—Closes the current experiment, standard Windows behavior. • Save—Saves the current experiment, standard Windows behavior. • Save As—Saves the current experiment offering the option to use a different name, standard Windows behavior. • Save Condensed Data—Save Condensed Data • Print—Prints the current view. • Print Preview—Previews the current view with the option of printing the document. • Exit—Closes the Scaffold window.
Edit 	<ul style="list-style-type: none"> • Copy—For each View copies to the clipboard the first table appearing at the top of the View. From there, the User can paste it into a third-party program such as Excel or Microsoft Word. • Find—Opens a find dialog box that searches the first table present in the Current View • Edit FASTA Database...—See Edit FASTA Databases. • Edit Peptide Threshold...—See Custom Peptide Filters • Edit Go Terms Options...—See Edit GO Term Options • Bulk Operation...—Tag, show and hide proteins. Expand and collapse clusters. Options also available on right click. See Protein List. • Preferences...—See Preferences • Advanced Preferences...—See Advanced Preferences

Chapter 4

The Scaffold Window

Menu	Menu Commands
View 	<ul style="list-style-type: none"> • Navigation Pane—Toggles the view of the Navigation pane. • Switch Sample View—Selects level of summarization in Samples View, see MS Sample vs BioSample summarization levels • Switch Display Options—See Display Options • Show Entire Protein Clusters—See Clusters in the Samples Table • Show Lower Scoring Matches—See Show Lower Scoring Matches • Show <5% Probabilities—See Show <5% Probabilities • Show Sample Notes—Toggles the view of the Information Panes • Show Hidden Proteins—Toggles the view of Hidden Proteins • Show GO Annotations—Toggles the view of GO Annotations Tab • Navigate—Navigates through tabs when present in a dialog or pane.
Experiment 	<ul style="list-style-type: none"> • Edit Experiment—See Edit Experiment • Edit BioSample—See Edit BioSample • Add BioSample—Initializes The Loading Wizard • Delete BioSample—Deletes a loaded biosample from the current Scaffold experiment. Particularly useful in the Load Data view. • Queue Files for Loading—See Queue Files for Loading. • Queue Files From Mascot Server For Loading—See Queue Files From Mascot Server for Loading.... • Queue Structured Directories For Loading—See Queue Structured Directory for Loading. • Apply New Database—See Apply New Database • Apply Protein Annotation Preferences—See Apply Protein annotation Preferences • Load and Analyze Queue—Available only when there are files listed in the loading Queue in the The Load Data View waiting to be loaded in Scaffold. When selected it opens the Load and Analyze Data page of the Loading Wizard. • Reset Peptide Validation—See Reset Peptide Validation • Apply Go Terms/Configure Go Annotations Sources—Applies imported annotations to the Samples Table. To import GO databases see Edit GO Term Options • Quantitative Analysis—See Quantitative Analysis...

Menu	Menu Commands
Export 	<ul style="list-style-type: none"> • Subset DATABASE— See Subset Database • Spectra— See Spectra • ProtXML— See ProtXML report • mzIdentML — See mzIdentML • Scaffold Batch— See ScaffoldBatch... • Scaffold Batch Archive— See ScaffoldBatch Archive... • Export To Excel: <ul style="list-style-type: none"> • Publication Report— See Publication report • Samples Report—Generates a tab-delimited Samples table appearing in the Samples View, see Samples report • Spectrum Reports— See Spectrum report • Peptide Reports—Generates a tab-delimited Peptide table for all proteins appearing in the Samples View, see Peptide report • Protein Reports—Opens the SQL dialog box see Protein report • Current View— See Current View report • Complete— See Complete report
Quant 	<ul style="list-style-type: none"> • Launch Q+ Quantitation Browser— When using Scaffold Q+ or Scaffold Q+S this command is available for switching to the Q+/Q+S quantitation window. • Edit Quantitative method/purity Correction— See Edit Quantitative Samples
Window 	<ul style="list-style-type: none"> • The User can use this menu to switch Scaffold views. Equivalent to clicking the buttons located in the Navigation pane

Menu	Menu Commands
Help 	<ul style="list-style-type: none"> • Help on Current View...—Opens the Online Help that is specific for the currently displayed topic. • Help Contents—Opens the Contents page for the Online Help. • Scaffold User's Guide—Opens the current Scaffold User's Guide. • Scaffold Q+ User's Guide—Opens the current Scaffold Q+S User's Guide. • Open Demo Files—Opens the folder where Scaffold demo files are stored. The User can choose any of the pre-loaded files to test Scaffold capabilities. • Scaffold FAQs/Resource Center—Opens the User's default web browser to the Home page of the Proteome Software resource center. • Show Log Files—Opens a folder containing Scaffold error_log and output_log files • Referencing Scaffold— See Referencing Scaffold • About Scaffold—Provides the release information for the current version of Scaffold, license information, contact information for Proteome Software Inc.. It also reports information about the system where Scaffold is installed, the amount of memory available to the software and the percentage of memory used by the application.
IdentityE	<ul style="list-style-type: none"> • Quantitation Options—Quantitation Option • Export IdentityE report—Generates a tab-delimited report containing the list of peptides used to calculate the intensities assigned to each protein in the list of identified proteins shown in the Samples view.

Merge

The command **File > Merge** allows the User to combine different Scaffold experiments into one single *.SF3 file. It is active only when an existing Scaffold experiment has already been created or opened. Selecting this command calls the **Import Scaffold File** file chooser from where it is possible to navigate to a *.SF3 file to be merged with the current opened Scaffold experiment. Once a file is selected the [Queue Scaffold Files for Merging](#) window opens allowing the User to add more files to the list of *.SF3 files to be merged.

When merging the different Scaffold experiments appearing in the list, the BioSamples included in each of them load into separate samples. If they happen to be equally named a number is appended at the end of the original denomination to distinguish them.

Queue Scaffold Files for Merging

After selecting the first file to be merged, the dialog **Queue Scaffold Files for Merging** appears. The function of the dialog is to help the User compile a list of files to be merged together in the same experiment.

The button **Add More Files** opens a file chooser which allows the User to locate, select and add files to the list appearing in the dialog. The **Merge** button merges the list of files to the original Scaffold experiment creating one or more new BioSamples for each file in the list. This means that if a merged file contained more than one BioSample the different BioSamples appear in the merged experiment.

Caution: It is not possible to delete a specific file from this list.

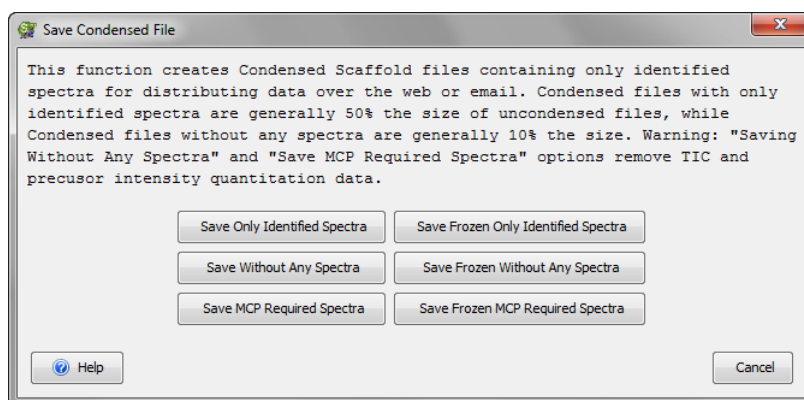
Once the files are merged the Delete Biological Samples operation can be used to delete any undesired BioSample.

Save Condensed Data

The **File > Save Condensed Data** menu option reduces the size of the *.SF3 file Scaffold saves.

Caution: Executing the Save Condensed Data changes the data in the running copy of Scaffold. Once the User saves condensed there is no undo that will restore all the data.

Figure 4-4:



There are six options for condensing data while it is saved:

- **Save Only Identified Spectra** —This option saves all the data that can be seen in the Scaffold Viewer. It does not save the spectra that were not matched to peptides. Saving with this option generally cuts the size of the *.SF3 file in half.
- **Save Frozen Only Identified Spectra** — This command condenses the saved output file just like the **Save Only Identified Spectra** option does except it also freezes the data in the files.
- **Save Without Any Spectra** —This option saves all the peptides and their scores but does not save any of the spectra.
- **Save Frozen Without Any Spectra** —This command condenses the saved output file just like the **Save Without Any Spectra** option does except it also freezes the data in the file.
- **Save MCP Required Spectra** — This saves only those spectra required by the proteomics journal.
- **Save Frozen MCP Required Spectra** —This command condenses the saved output file just like the **Save MCP Required Spectra** option does except it also freezes the data in the file.

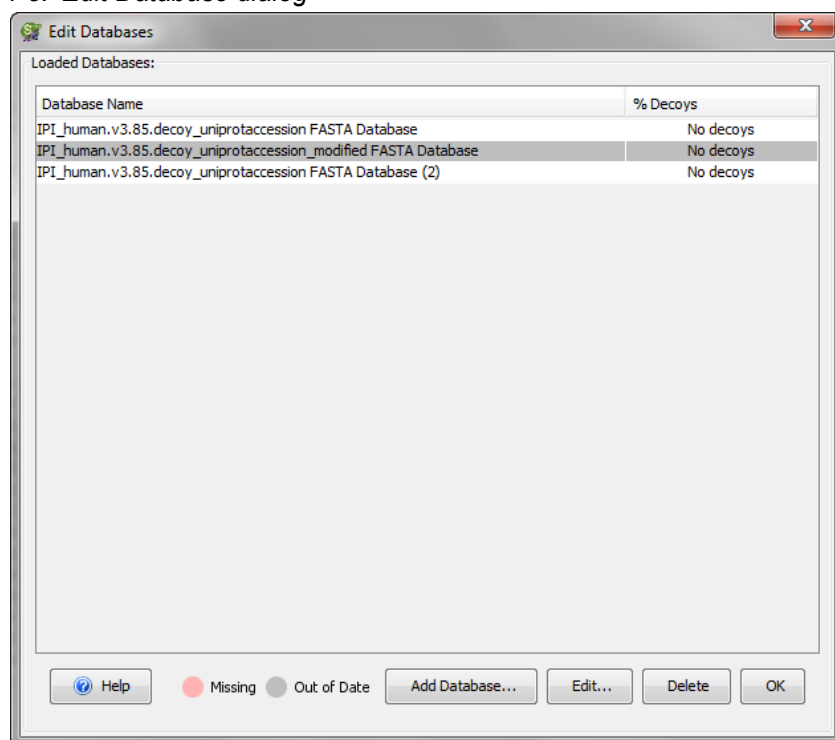
Since the spectra are 90% of the bulk of the data, an *.SF3 file saved without spectra will be reduced to only about 10% of the size of the uncondensed file.

FASTA databases in Scaffold

Edit FASTA Databases

To add and parse databases the User should open the **Edit Databases** dialog either selecting the menu option **Edit > Edit FASTA Databases** or clicking the button **Add New Database** located in the Search Database pane in the [Load and Analyze Data](#) page in the Scaffold loading Wizard. The selection opens the **Edit Databases** dialog which contains a table listing the databases already available in Scaffold and a number of functional buttons.

Figure 4-5: Edit Database dialog



Loaded Databases table

This table lists all the databases already available in Scaffold with information about the percent of decoys included in each of them. A database in the list might appear highlighted in various colors as a warning.

- Pink highlight: *Missing database*. Scaffold cannot connect with the database using the current stored information.
- Grey highlight: *Out of date database*. The database indexing is not updated with the Scaffold version in use.

When selecting a highlighted database the button **FIX** appears. Depending on the issue,

clicking fix either calls the parsing method directly to build or rebuild the related index file or asks for a location where to find the database if it was moved.

Add Database button

This button adds a new database to Scaffold. When selected a file browser appears allowing the User to point Scaffold to the location where the database he/she wants to load is stored. Once the FASTA file is selected the [Parsing Method Dialog](#) appears.

Edit button

To edit one of the existing databases the User has to select a name from the Loaded Databases table and click **Edit**. The [Parsing Method Dialog](#) appears.

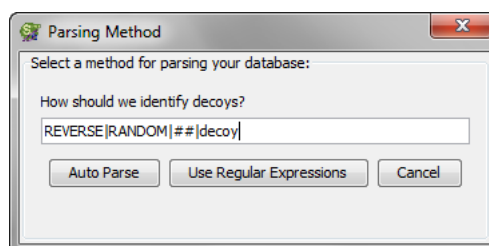
Delete button

To delete one of the existing databases the User has to select a name from the Loaded Databases table and click **Delete**.

After parsing rules are applied or databases are deleted the User can click **OK**.

Parsing Method Dialog

Figure 4-6: Database Parsing methods dialog



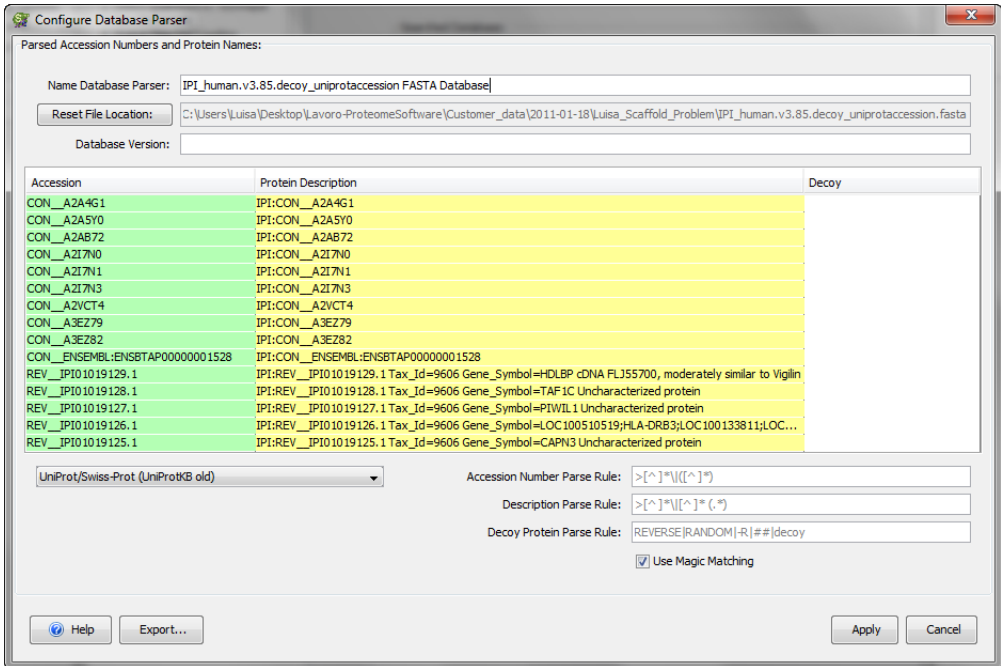
This dialog Box allows the user to select one of the two parsing methods Scaffold uses to align protein names and accession numbers:

- **Auto Parse**— This option provides an automatic way of searching for the optimal accession numbers between the database and the type of data loaded into Scaffold. It initially identifies the type of parsing rule that better fits both the data and the selected database. It then matches the rule protein by protein while ensuring uniqueness. If a protein does not include the type of rule initially selected, **Auto Parse** looks for other rules more compatible with the specific protein accession number and defaults to a more general accession number if everything fails.
- **Use Regular Expressions**— This options opens the [Configure Database Parser](#) window which allows the verification of the protein names and accession numbers alignment for Scaffold's indexing and parsing. It also gives the possibility to modify the parsing rules according to the User's needs.
- **How should we identify Decoys?**— This box contains typical tags used to label decoy proteins in a database. When parsing a database that contains decoys the User should

make sure that the decoy identification tag used in the his/her database is included in the list.

Configure Database Parser

Figure 4-7: Configure Database Parser dialog



This dialog contains tools to help the User describe and edit the location of the selected database:

- **Name Database Parser**—Through this text box the User can change the name assigned to the database when loaded
- **Reset file location**— By clicking this button the User can point Scaffold to a different location where the database is stored.
- **Database version**— This text box can be used to define the Database version

The dialog also contains tools to help the User parse the database as needed:

- **Parsed Accession numbers and protein names table** — This table lists a sample of the protein accession numbers and descriptions contained in the database. The list includes proteins selected from the top and the bottom of the database file to give an idea of the type of accession numbers used in the database. The accession numbers and protein descriptions are shown parsed according to the rules selected from the pull down list of parsing rules, see below.
- **Parsing Rules Pull down list**— The list includes a number of standard parsing methods for different types of databases and their related accession numbers format (like Swiss-prot, Uniprot-sprot etc). Once a particular rule is selected the different related parsing

strings are shown in the test boxes located on the right hand side of the list. The User's specified selection in the list allows editing of the rules appearing in the text boxes and when clicking in a different text box Scaffold automatically verifies the validity of the inputted rule.

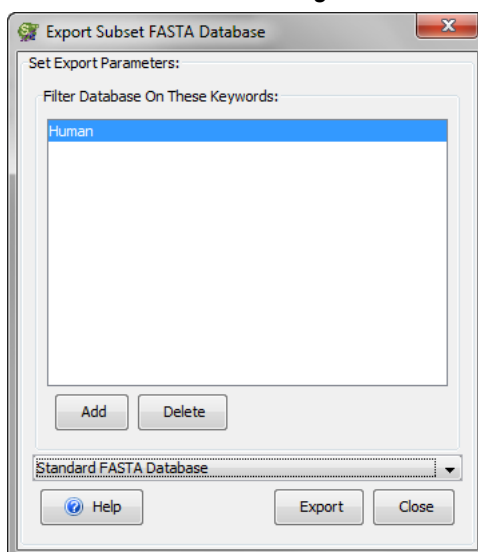
- **Magic Matching check box**- This tool optimizes the accession numbers available for the proteins to better match databases and loaded data. Once a parsing rule has been selected by the User through the pull down list, Magic Matching checks protein by protein if that type of accession number properly matches the protein and finds alternatives if it does not. If no alternatives are available it defaults to a generic accession number.

In the bottom left corner of the dialog there are two buttons one that calls for the Online-help and another one, the **Export** button, that calls the [Export Subset FASTA Database](#) dialog to create a subset database or a decoy database.

Export Subset FASTA Database

This dialog is called by the **Export** button located in the [Configure Database Parser](#) dialog. It provides tools to create a new filtered FASTA database or decoy database starting from the one selected in the Configure Database Parser dialog. It contains a list and pull down menu.

Figure 4-8: Export Subset FASTA Databases dialog



- **List of Filter Keywords** - Any of the keywords in the list is used to filter the original database for accession numbers that contains them. Note that the keywords are not case sensitive, do not have to be complete words, and can be multiple word phrases. Key words can be added to and deleted from the list using the buttons present at the bottom of the list. The button Add opens the Add Keywords Filter dialog where the User can type in a new word.

This option is most often used to create a FASTA file for a specific species from a huge database. The taxonomy of the protein is listed in different ways in different databases,

so the User needs to choose keywords appropriately. For example, to select only bovine proteins from the complete UniPROT database, enter the keyword "_BOVIN" or to select rat proteins, enter the keyword "_Rattus".

- **Database type pull down menu-** It shows the list of possible types of databases that can be created through this function:
 - *Standard FASTA database* - This option is used when the User wants to filter a large database with specific keyword to reduce its size.
 - *Reverse FASTA Database* - Each accession number has a "-R" appended to it. The protein description is unchanged. The protein sequence is reversed.
 - *Random FASTA Database* - Each accession number has a "-R" appended to it. The protein description is unchanged. The protein sequence is scrambled in a random manner.
 - *Reverse Concatenated FASTA Database* - Each protein in the original FASTA file appears unchanged, but it is preceded in the FASTA file by the reverse protein ("-R" appended to accession number and sequence reversed). This database is twice as long as the original.
 - *Random Concatenated FASTA Database* - Each protein in the original FASTA file appears unchanged, but it is preceded in the FASTA file by the randomly scrambled protein ("-R" appended to accession number and sequence scrambled). This database is twice as long as the original.

After selecting the appropriate options the User can click Export to save the new database.

Edit GO Term Options

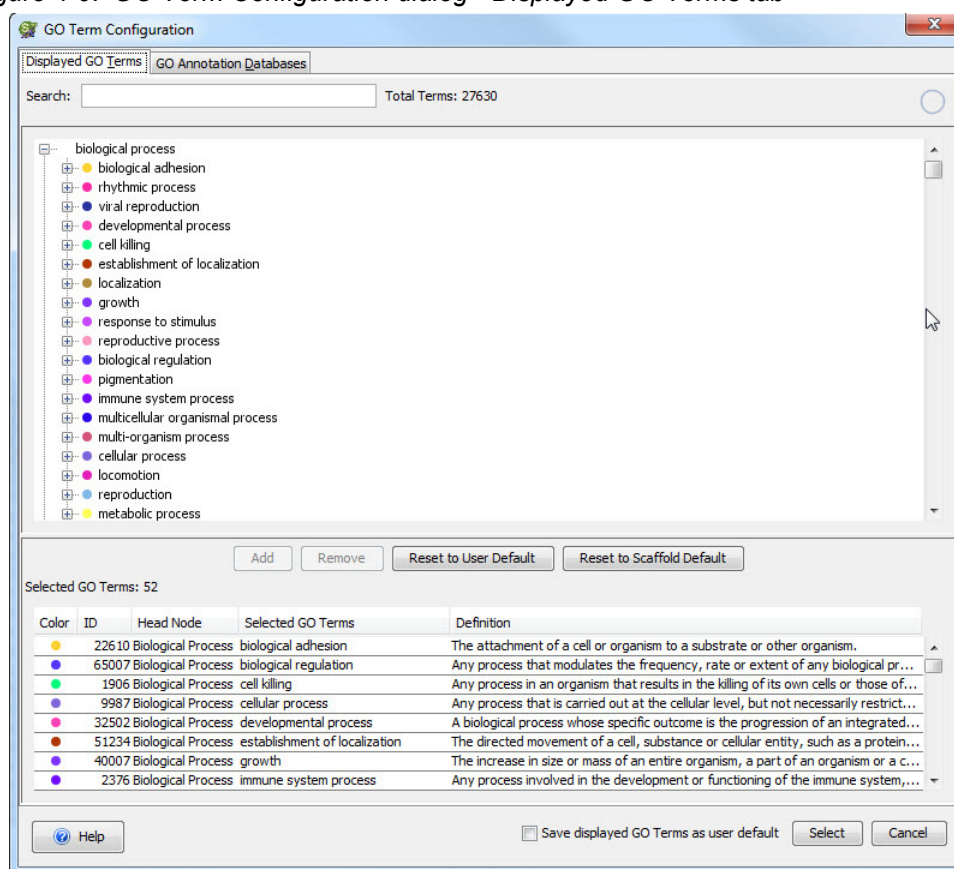
When going to the main menu and selecting **Edit > Edit GO Term Options**, the **GO Term Configuration** dialog opens.

It contains the following tabs:

- [The Displayed GO Terms Tab](#)
- [GO Annotations Tab](#)

The Displayed GO Terms Tab

Figure 4-9: GO Term Configuration dialog - Displayed GO Terms tab



Through this tab the User can create and modify a custom list of GO terms. The list is then displayed as extra columns in [The Samples Table](#) whenever the terms are present in the experiment.

The Tab is divided into sections:

- **Search Field** - Searches terms available in the GO terms database loaded in Scaffold.
- **GO Tree list** - Hierarchical list of all the terms present in the loaded GO database

Chapter 4

The Scaffold Window

- **Add and Remove GO terms** - Provides tools for creating the custom Display list
- **Display List** - List of GO terms selected by the User that will be visible in [The Samples Table](#).
- **Save and Apply**- Allows the User to save the current Display List if changed

To create a new custom GO terms Display List the User needs to follow these instructions:

1. If the **Display List** is not empty select all the rows and press delete.
2. Search and select any GO term of interest present in the loaded GO database either by typing a name in the **Search Field** or by selecting a row in the **GO Tree List**.
3. Click **Add**; the selected term or group of terms is added to the **Display List**. Terms may be selected individually or by domain or group. If a group or domain is selected, all terms in that group will be added to the **Display List**.
4. To remove terms from the **Display List**, select a term or group of terms to be discarded then click **Remove**.
5. To save the current selections as User Defaults check the box **Save displayed GO terms as user default**.

When a Scaffold experiment is saved, the displayed GO terms are saved within the *.SF3 file.

When a new file is created, or when Scaffold is closed, the list of displayed GO terms is unchanged. To reset the list to the defaults, the user may click the **Reset to User Default** or the **Reset to Scaffold Default** button.

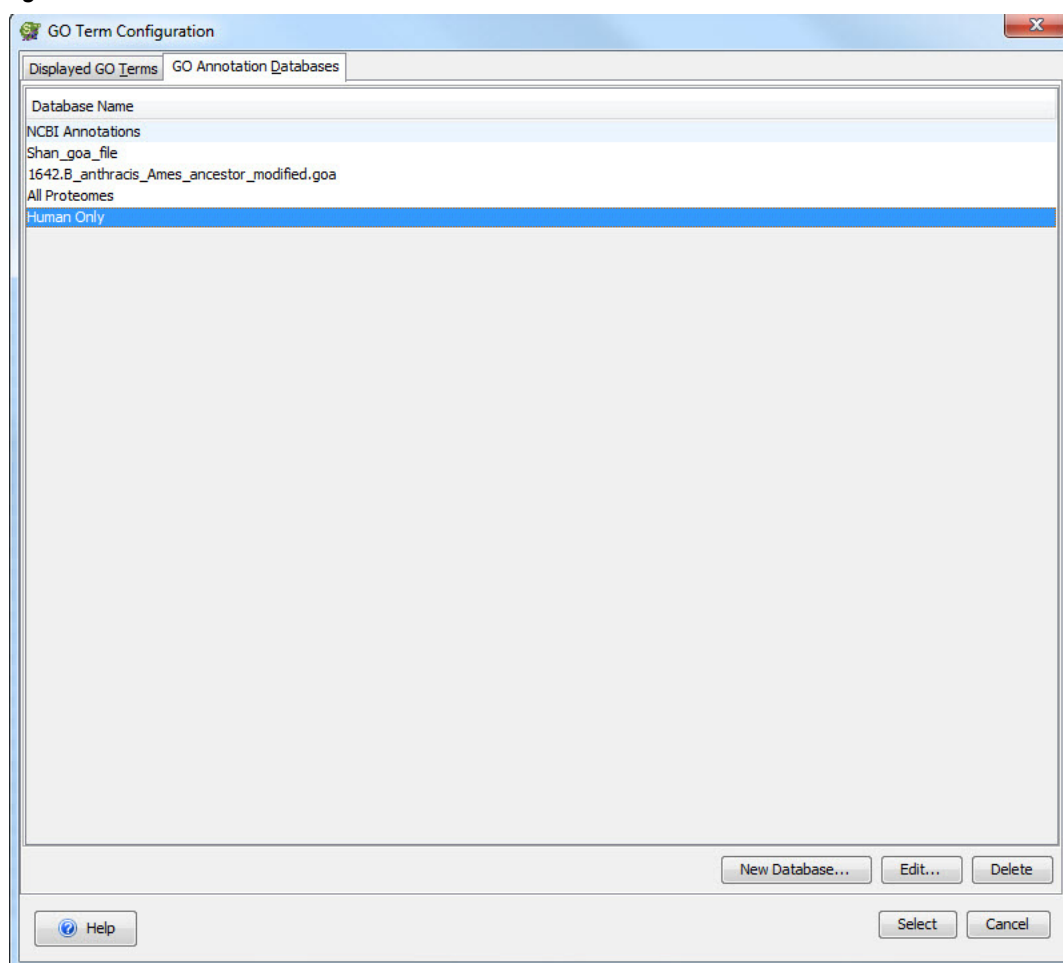
GO Annotations Tab

This tab contains a table which lists all the GO annotations databases already imported in Scaffold and the option NCBI Annotations. The User can populate the table with existing or custom created GO terms databases through the [Import annotations](#) function and then select among them which is the one he/she wants to use to annotate the protein list appearing in the Samples Table.

When NCBI Annotations is selected Scaffold queries the NCBI website through the INTERNET. This option is the only one available when Scaffold is initially installed and before the User imports GO databases on his/her own. Nevertheless it needs to be selected before being able to apply GO terms to the protein list.

The GO Annotations Tab also includes a search box and the [Import annotations](#) button to import GO databases.

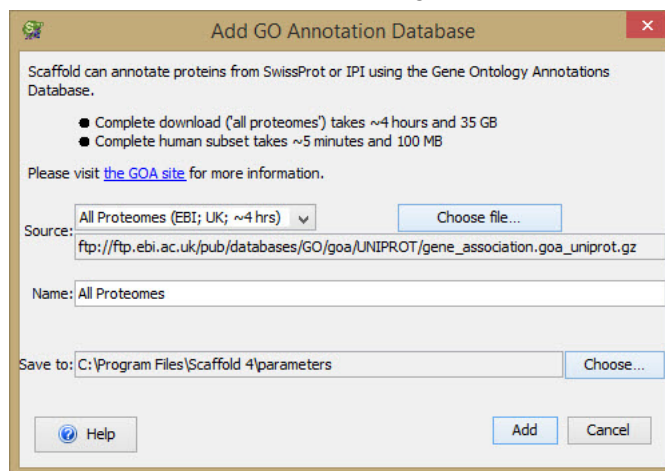
Figure 4-10: Go annotations tab



Import annotations

The **Import Annotations** button opens a dialog through which the User can import GO databases in Scaffold. A pull-down menu directs Scaffold to different locations where the GO Database can be downloaded.

Figure 4-11: Add GO Annotations Database dialog



The pull down list includes the following items:

- **All Proteomes** - provides a complete download of the unfiltered UNIPROT GO Database. It approximately takes 2 hours to download a 4 GB file.
- **Human Only** - provides a download of the human subset. It takes about 10 minutes to download.
- **Other Website** - the user can type in a website address from where a GO Database can be downloaded.
- **Other File** - the User can direct Scaffold to a location in his/her computer where the GO database is stored.

After one of the options is selected, clicking **Add** starts the operation of importing the GO annotation database into Scaffold. A new row appears in the list of already loaded databases showing the name of the newly added database and the number of annotations included in it.

After selecting the GO database of interest, clicking **OK** closes the dialog and Scaffold is now ready to annotate with GO terms the protein list in the Samples table. The User can start the process by choosing, the now available option, **Experiment > Apply GO Terms**.



*The command **Experiment > Apply GO Terms** is available for use only when one or more GO Annotations databases are loaded into Scaffold.*

Preferences

The Preferences dialog provides a series of modifiable options organized in a number of different tabs. Through this dialog the User can modify parameters and settings to customize the way Scaffold experiments appear and run.

Selecting the menu item **Edit > Preferences** opens the **Preferences** dialog which contains the following tabs:

- [Internet](#)
- [Memory](#)
- [Processors](#)
- [Web Link](#)
- [Mascot Server](#)
- [Display Settings](#)
- [Password](#)
- [Paths](#)

Internet

In the Internet Settings dialog the User can enter a Proxy server name or IP address and a proxy port number. Through check boxes in this dialog box, the User may:

- **Allow Scaffold to connect to the Internet** If this box is unchecked, then Scaffold cannot access the Internet. Users may want to have this box unchecked if their organization prevents connections to the Internet.
- **Use HTTP Proxy Server**
 - **Proxy Server name (or IP address)**
 - **Proxy port number**

Proxy servers may be used by an organization's IT departments to filter communications to and from the Internet. If that is the case, then Users need to set the Proxy Server Name and Port Number. Users can check if there is any need to use proxy server settings by looking at how their web browser is connected to the web.

Memory

This tab allows the User to set the maximum amount of memory that Scaffold is allowed to use. Scaffold is a memory-intensive program and needs a large amount of RAM to be able run at a decent speed.

When setting the amount of memory Scaffold should use it is important leave enough memory for other programs to run.



- *The new memory setting will take effect only after the application has been closed and restarted.*

Processors

This tab provides information to Scaffold about the maximum number of processors available for threading computations. The default value is the maximum number of processors available in the system where the application is installed.

The Scaffold application in itself uses only two threads. Assigning more than two threads to Scaffold mainly affects how fast the X! Tandem version bundled with Scaffold executes thus optimizing the throughput.

Web Link

Through this tab the User can add, change or delete the on-line protein lookup databases links. These web sites appear in the **Lookup Accession Number in:** pull down list found in the [Protein Information pane](#) in the Samples View.

When selecting these databases it is important to note that they do not need to be the same as the FASTA database that was used in the searches, but they must have the same type of accession numbers.

Clicking New Database opens the **Configure Web Link** dialog where information for a new Online database can be added.

The linked database can be either a public database, or an internal one. It just has to have a URL that queries the database. This link could also be to a web site that performs a calculation or does a BLAST.

The User might want to set up several links to the same database that query it in using different types of accession numbers.

The **Edit** button allows the User to modify a web link to adjust the URL if it has changed, or a better URL has become available.

Mascot Server

Scaffold can load data directly from a Mascot Server. This tab contains a text box where the User can set up the connection to the server by writing the web address of the available Mascot server.

The button **Test Connection**, located on the lower right corner of the tab page, provides a quick way to check if the connection works properly.

- If no security is implemented, Scaffold connects directly to the Mascot sever. When the **Test Connection** button is clicked, a message appears stating whether the connection was successful or not.

- If security is enabled on the Mascot Server and the **Test Connection** button is clicked a login window pops up asking for an account name and password. The User has to make sure that the account he/she is using in Mascot has administrative rights.



Scaffold does not download files from a Mascot Server if the User is logged on as a GUEST and an error is shown.

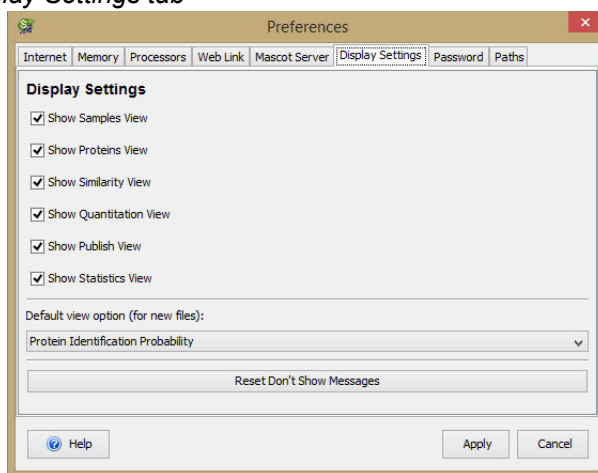
Display Settings

Scaffold provides different ways to look at the data included in an experiment through different views: Load Data, Samples, Proteins, Similarity, Quantify, Publish and Statistics. The Display Settings tab allows the User to decide which of the available Views is visible.

Through this tab the User can also define which default Display Options is selected when a new experiment is created and reset messages that were selected not to show anymore, like the initial repetitive dialogs that appear when the Wizard is opened.

The tab includes a check box list of all the views available in Scaffold, a pull down list of the Display Options to select the appropriate default value and a reset button.

Figure 4-12: Display Settings tab



Check box list

If one of the views in the list is not checked it will not display and the corresponding button in the Navigation pane will also not be visible.

These settings are saved with the experiment when the *.SF3 file is created. For example, if the User turns off the Statistics and Proteins views and saves the file, when the file is reopened, only the Load Data, Samples and Publish views are visible.

This feature can be useful when sending the results to someone who doesn't need to see certain details or to check the validity of the statistical analysis.

Access to Display Settings may be controlled by a password by checking the appropriate box

Chapter 4

The Scaffold Window

on the **Password** tab of the Preferences dialog. This means that the User can control which pages his/her collaborators can view. The User can use this in conjunction with the password protection on the filters to control which proteins can be viewed.

Pull down list

The bottom portion of the window stores the preferred Display Options for the Samples View when a new experiment is created. A pull down menu will allow the user to select what information Scaffold will initially show when the loading phase is completed and the Samples view is initially shown.

Reset button

The **Reset Don't Show Messages** button restores the messages that were checked to not show again when requested.

Password

Through this tab the User can select to use a password to protect certain views and operations available in a Scaffold experiment once it is saved in a *.SF3 file. For example, the User can set filter thresholds to display only data with above 90% confidence or restrict access to only the Samples and Publish pages, in this way hiding the messy details on the Proteins and Statistics pages. The User can also prevent anyone from reanalyzing his or her data by locking the export of the spectra.

A password gives the User control, control of what the people viewing data can see and do.

- **Use Password...** - Turns on and off the password protection
- **Protect Exporting Spectra** - Password required to export spectra.
- **Protect Resetting Thresholds** - Password required to change Min Peptide filter or define custom filters.
- **Protect Changing Display Settings** - Password required to hide or display hidden pages.
- **Protect Hidden Proteins** - Password required to access **View > Show Hidden Proteins**

Paths

The Scaffold Installation comes with a generic UNIMOD database typically used to unify modifications naming among different search engines when their results are loaded in the same Scaffold experiment.

This Tab allows the User to select alternative UNIMOD databases when loading data:

- **Do not use UNIMOD** - This option tells Scaffold to retrieve the information about modifications directly from the search engine results that are being loaded.
- **Use Scaffold default UNIMOD** - This options tells Scaffold to use the default UNIMOD database.

- **Use a custom UNIMOD file** - This option allows the User to direct Scaffold to a location where a custom UNIMOD database is available and retrieve the modification information from the selected database.

Advanced Preferences

To have Scaffold correctly compute the discriminant score used by the PeptideProphet algorithm, see [Keller \(2002\)](#), the loaded data files need to contain specific pieces of information. Depending on how the search engine parameters are set, it might happen that some vital information used to properly calculate the discriminant score is discarded and not saved in the output files. When this happens, the peptide probability assignments, and consequently the protein probabilities, are computed by Scaffold in an unreliable fashion.

To compensate for this problem the Advanced Preferences Edit menu selection opens a dialog where the User can choose which scoring function is used by Scaffold in analyzing the search engine data using PeptideProphet.

The dialog presents separate tabs for the Sequest and Mascot search engines:

- [Sequest tab](#)
- [Mascot tab](#)

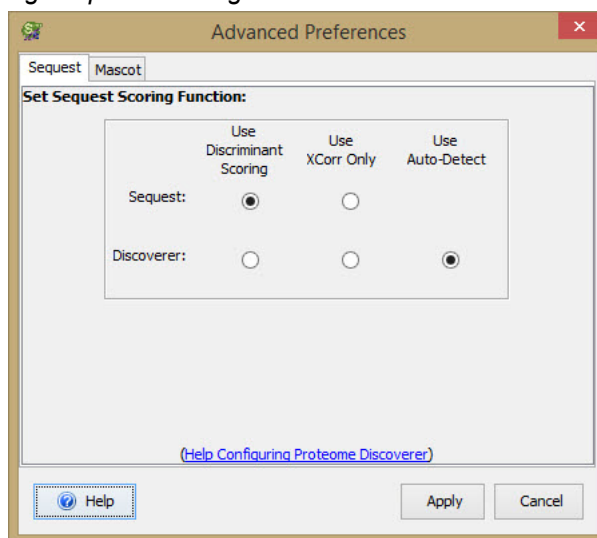
Sequest tab

When running Sequest searches using Proteome Discoverer and using the default settings suggested by Thermo, the *.MSF output files do not contain records of any unassigned spectra. This missing information directly affects the ability of Scaffold to calculate the delta Cn score which is included in the formula used by Scaffold to compute the Sequest discriminant score in the PeptideProphet algorithm. This formula is a normalized version of the Sequest XCORR score and depends upon the charge state of a peptide. For example, for charge +2 the discriminant function is $8.36 * XCORR + 7.39 * DeltaCn - 0.19 * \ln SpRank - 0.31 * deltaMass - 0.96$

For PD version 1.3 and above, we were able to identify the Sequest parameters that affected the amount of information recorded in the *.MSF files once the search is ended. Proper suggestions concerning their adjustment are recorded in [Configuring Proteome Discoverer Sequest and Mascot](#).

Unfortunately, those parameters cannot be adjusted in PD version 1.2. To address this issue, we created the Sequest tab in the Advanced Preferences where the User can specify what type of scoring function Scaffold uses in the PeptideProphet algorithm when loading Sequest searches.

Figure 4-13: Setting Sequest Scoring Function



The Sequest tab includes a table with the different available options:

- **Generic Sequest** - The User can select either Discriminant Score or XCorr Only
- **Discoverer Sequest** - The User can select Discriminant Score, XCorr Only or use Scaffold auto-detect function that checks, in the loading phase, if all the needed information is included in the data files. When all the proper information is present Scaffold uses the Sequest discriminant score; otherwise XCorr Only.

Note: When XCorr Only is used the list of identified proteins will be shorter, due to the fact that XCorr Only reflects more stringent conditions when calculating the peptide probabilities.

Mascot tab

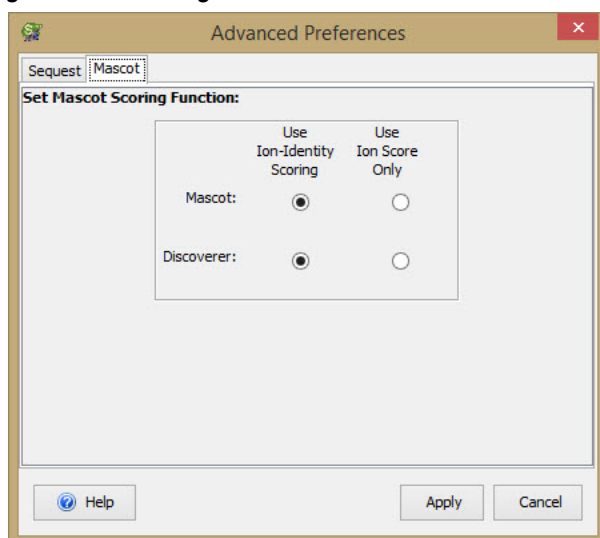
To assure a proper definition of the Mascot scoring function used by Scaffold to calculate the peptide probabilities with PeptideProphet, a certain amount of information needs to be available in the output files created by Mascot.

The Mascot scoring function Scaffold normally uses is the Mascot Ion Score minus the Identity Score. The Identity Score Scaffold uses is the level that has a 5% probability of being due to a random match. Mascot's concept of probability is somewhat different than Scaffold's, but roughly speaking if you set Scaffold's Min Peptide probability to 95%, the black vertical line on the Mascot Histogram should be close to zero on the discriminant scale.

Depending on the parameters set for the searches, at times a reduced amount of information is exported to the output files and pieces of information needed to calculate the Identity Score is missing. This affects the values of the calculated peptide probabilities and consequently the probability assigned to the list of identified proteins.

When this happens through the Advanced Preferences Mascot tab the User can select Ion Score Only as the scoring option used by PeptideProphet, reducing in this way the error created by the improper calculation of the Ion-Identity Scoring.

Figure 4-14: Setting Mascot Scoring Function



The Mascot tab includes a table listing the different programs producing Mascot search results and radial buttons for selecting which scoring function Scaffold uses in the PeptideProphet algorithm:

- **Generic Mascot** - The User can select either Ion-Identity Scoring
- **Discoverer Mascot** - The User can select either Ion-Identity Scoring

For PD version 1.3 and above, we were able to identify the Mascot parameters that affected the amount of information recorded in the *.MSF files once the search is ended. Proper suggestions concerning their adjustment are recorded in [Configuring Proteome Discoverer Sequest and Mascot](#).

Note: Note that selecting **Use Ion Score Only** provides a list of proteins different in length than when **Use Ion-Identity Scoring** is selected.

Configuring Proteome Discoverer Sequest and Mascot

When searching MS data using Thermo Proteome Discoverer (PD), it is possible to adjust the amount and type of information stored in the output files, *.MFS files. Some of the settings used to reduce the quantity of stored information in the output files, have default values that discard most of the low hit spectra. When loading this types of *.MSF files in Scaffold the User might encounter inconsistencies in the way Scaffold assigns probabilities to the peptides and proteins. To be able to properly run the different scoring algorithms, like the LFDR and the PeptideProphet algorithms, in a statistically meaningful way, Scaffold needs a certain number of false hit spectra included in the imported data.

With *.MSF files created using the default settings in PD 1.4 and older, it is quite clear that many of the false hit spectra are discarded and so are not saved in the search result files. Once the files are loaded into Scaffold, the discriminant score histogram displayed in the Scaffold Statistical view looks quite skewed and the related calculated protein probabilities become unreliable.

To address this problem, we introduced a set of

The [Advanced Preferences](#) in Scaffold provide tools to deal with this issue. Furthermore an auto-detect feature, that comes into play when *.MSF files are loaded into Scaffold, selects which of the option in the Advanced Preferences best suits the data that is being loaded.

We also recommend the User to adjust the Advanced Options available in the latest version of PD to allow a less stringent selection of the spectra saved in the MSF files as described in:

- [PD Sequest suggested Settings](#)
- [PD Mascot suggested settings](#)
- [PD Sequest HT suggested settings](#)

PD Sequest suggested Settings

In Proteome Discoverer 1.4 and older the Work Flow settings for Sequest include parameters that determine the amount and type of Spectra saved in the *.MSF output files. Those parameters are located in the **1.1 Peptide Scoring Options** section visible only when Show Advanced Parameters is selected.

We advise the User to adjust the following parameters to their minimum value:

- Absolute XCorr Threshold = 0
- Fragment Ion Cutoff Percent=0
- Peptide Without Protein XCorr Threshold = 0

so that Scaffold is able to find the information needed for proper DeltaCn calculations. If the PD 1.4 and older User chooses to adjust the peptide Scoring Options as we described above, the *.MSF file created retains the lower scoring matches and can then be loaded in Scaffold using the regular discriminant score either specifically selecting the option in Scaffold [Advanced Preferences](#) or selecting auto-detect.

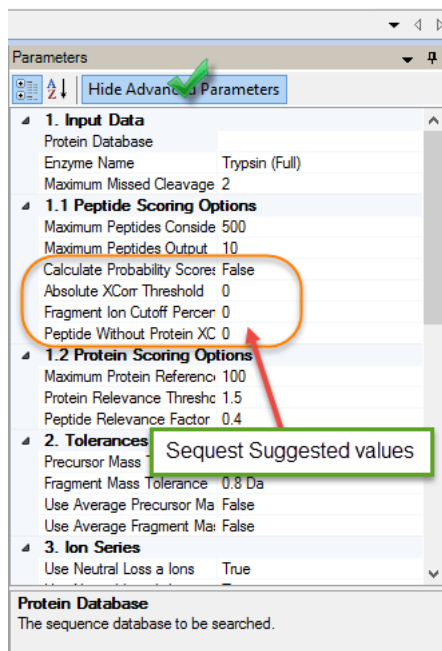


The PD options shown below are not available in PD 1.2. In this case, selecting the auto-detect feature will ensure the proper handling of the data.

Chapter 4

The Scaffold Window

Figure 4-15: Sequest Advanced Parameters Peptide Scoring Options in Proteome Discoverer



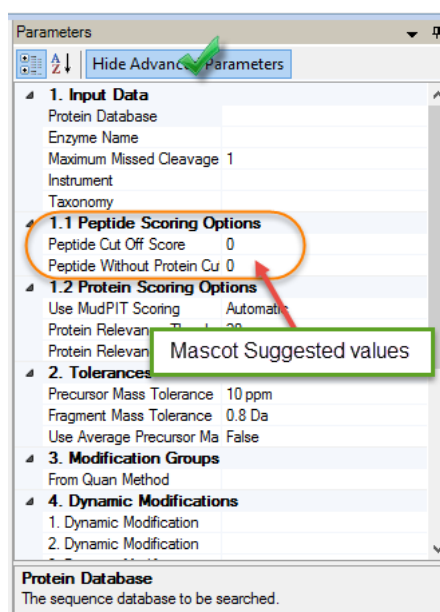
PD Mascot suggested settings

In Proteome Discoverer 1.4 and older the Work Flow settings for Mascot include parameters that determine the amount and type of Spectra saved in the *.MSF output files. Those parameters are located in the **1.1 Peptide Scoring Options** section visible only when Show Advanced Parameters is selected.

We advise the User to adjust the following parameters to their minimum value:

- Peptide Cut Off Score = 0
- Peptide Without Cut Off Score = 0

Figure 4-16: Mascot Advanced Parameters Peptide Scoring Options in Proteome Discoverer



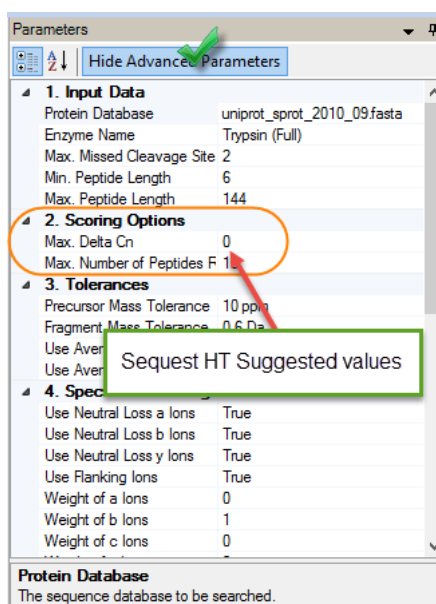
PD Sequest HT suggested settings

In Proteome Discoverer 1.4 there is a new version of Sequest available, called Sequest HT. As for regular Sequest and Mascot, the Work Flow settings for Sequest HT include a parameter that determines the amount and type of Spectra saved in the *.MSF output files. That parameter is located in the **2. Scoring Options** section.

We advise the User to adjust the following parameter to its minimum value:

- Max. Delta Cn= 0

Figure 4-17: Sequest HT suggested Scoring Options in PD 1.4 and higher



Show Lower Scoring Matches

The command **View > Show Lower Scoring Matches** toggles the option of rendering visible in the Samples Table the presence of a protein in a sample even if it does not meet the current filters and thresholds.

In some cases several samples may identify a protein at very different confidence levels. For example, sample 1 may identify protein A with 95% probability and sample 2 may only identify it with 60% probability.

- If the option **View > Show Lower Scoring Matches** is selected, then the filters and thresholds affect only which protein rows are shown and both the 95% and the 60% values would be displayed, even if the protein threshold was set at 90%.
- If the **View > Show Lower Scoring Matches** option is not selected, then the sample values that do not meet the filter values are suppressed. This means that the 95% value for sample 1 would be shown, but no value would be shown for sample 2.



*It is particularly important to be aware of the status of the **View > Show Lower Scoring Matches** option since it affects the counts shown in the table and varies the quantitative values.*

Note: When probabilities are lower than 5% values will not be shown unless [Show <5% Probabilities](#) is selected.

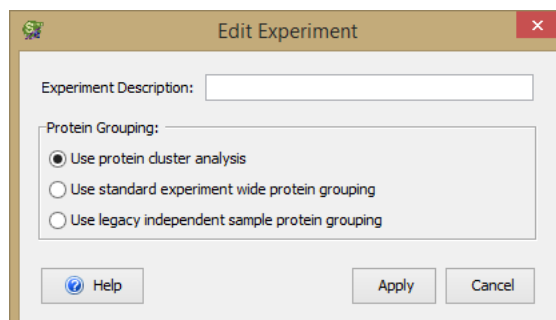
Show <5% Probabilities

When Show lower Scoring Matches is not selected values for proteins that have probabilities less than 5% are not shown. This option allows those values to be visible.

Edit Experiment

The menu option **Experiment > Edit Experiment** opens a dialog where the User can add or edit a description of the experiment.

Figure 4-18: Edit Experiment menu option



In the full version of Scaffold this dialog also contains the Protein Grouping pane where it is possible to toggle the various protein grouping options available in the program.

- **Use protein cluster analysis** - When selected Scaffold uses the [Shared Peptide Grouping and Protein Cluster Analysis](#) to group and pair the list of identified proteins.
- **Use standard experiment wide protein grouping** - When selected Scaffold uses the [Legacy Protein Grouping](#) with no clustering.
- **Use legacy independent sample protein grouping** - Scaffold uses the [Legacy Protein Grouping](#) with no clustering, but the grouping is done within biosamples and not across biosamples.

Edit BioSample

The menu option **Experiment > BioSample** opens a dialog where the User can add or edit the name of the sample, its category and description, see [Organize Samples In Categories](#).

Figure 4-19: Edit BioSample

Note that, defining these parameters in a concise and consistent matter is quite useful since Scaffold uses the Sample and Category names in sorting columns in the Samples View.

The dialog also shows whether the data was loaded using the Mudpit or condensed data options.



- *While this option may be selected in any view, it is highly recommended to use it only from the Load Data view to facilitate the selection of the BioSample that is going to be modified.*
- *To avoid unintended inconsistencies in category names, choose the appropriate name from the drop-down list whenever available.*

Organize Samples In Categories

When BioSample are defined in the [The Loading Wizard](#), they can also be organized into Categories. If BioSamples are not originally defined in Categories, they can be organized later by selecting the menu option **Experiment > Edit Biological Sample**.

Categories are useful in two ways. The first is that the columns in the [The Samples Table](#) have all the BioSamples grouped into categories. For example if the samples are put into categories “Treated” and “Control”, then the samples in the “Control” category will be grouped together to the left of the samples in the “Treated” category.

The second way categories are useful is to organize the samples in order to find which proteins among categories are differentially expressed. Scaffold offers several options for comparing the expression level of each protein between categories. The quantitative analysis terms, **Experiment > Quantitative Analysis...**, T-Test and ANOVA both measure the statistical probability of difference between categories. Likewise the Quantify View organizes data in categories.

Apply New Database

Through the use of the menu option **Experiment > Apply New Database** the User can address the following situations:

- **Incorrect parsing of protein accession numbers** - When this happens, [The Samples Table](#) reports question marks in the Molecular weight column while in the Proteins View the protein sequence is missing. Most of the time the cause of this problem is related to an incorrect parsing of the database selected when loading the data into Scaffold. Either the database is not the same as the one used for the searches or Scaffold was not able to apply proper parsing rules to connect the accession numbers appearing in the search results to the database used when loading. Selecting another database or re-parsing the database used in the loading phase typically resolves the problem.
- **Loading data searched using multiple databases** - In this case data is typically loaded selecting one of the databases used for the search. Proteins identified with the other databases are not correctly parsed and their molecular weights appear as a question marks. This problem can be resolved by selecting and applying the other databases used in the analyses. Scaffold picks up the unidentified proteins and resolves the question marks appearing in the molecular weight column and retrieves the protein sequences appearing in the Peptides View.

When selected the menu option opens the Select Database dialog showing the list of FASTA databases currently loaded in Scaffold. The User can then choose a different database and apply it to the current list of proteins. The functionality present in this dialog are the same as those appearing in [Edit FASTA Databases](#).

Load and Analyze Queue

This command is available only when there are files present in the Loading Queue shown in the [The Load Data View](#). When selected it opens the [Load and Analyze Data](#) page of the Loading Wizard where the User can select the proper loading parameters and load the data in Scaffold

Reset Peptide Validation

In the Proteins View Scaffold provides tools to manually inspect the identification of peptides. A validation check box records the status of a peptide; when selected a peptide is considered valid. The User, upon visually inspecting the related spectrum, can invalidate a peptide by manually deselecting its check box.

The menu option **Experiment > Reset Peptide Validation** provides a global tool to automatically validate all peptides above a specified probability by selecting the related check box and unchecking those below it. When a different probability is selected the command resets all previous user validations. The default probability is 0%.

This function can be used in two possible scenarios of the peptide validation process:

- **Globally create a set of validated peptides based on probability assignments.**
Peptides are considered valid only if their probability is greater than the minimum amount set in the pull down menu **Minimum Peptide Probability**, its default probability being 0%. Select a different value and click Apply. All peptides with probability less than the Minimum Peptide Probability will be shown unchecked in the Proteins View and not considered for analysis.

- **Reset the manually validated peptides to their initial status**, the initial status being identified by the minimum peptide probability recorded in the pull down menu.

Apply GO Annotations/ Apply NCBI/ Configure GO annotation Sources

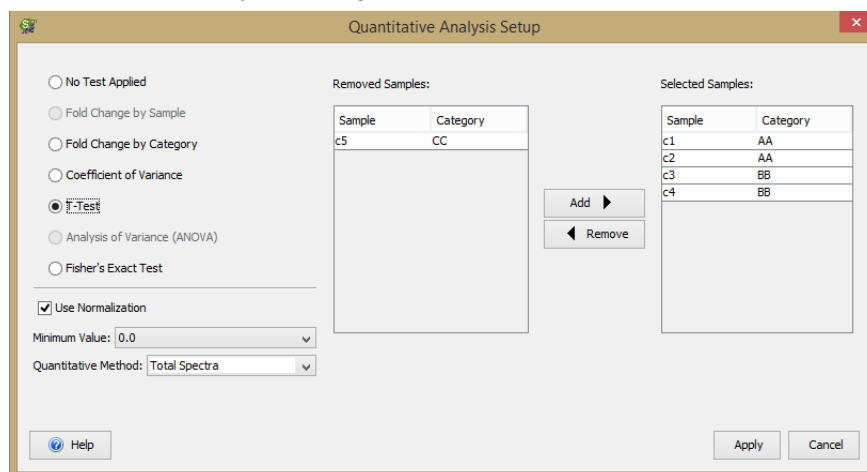
This menu command can have three possible statuses:

- **Apply GO Annotations** - This status appears when a GO annotations database has been imported and selected from the **Edit > Edit GO Terms Options...**, [GO Annotations Tab](#) tab.
- **Apply NCBI** - This status appears when NCBI Annotations has been selected from the **Edit > Edit GO Terms Options...**, [GO Annotations Tab](#) tab.
- **Configure GO annotation Sources** - This status appears when the User has yet to select a GO annotations database or NCBI Annotations from the **Edit > Edit GO Terms Options...**, [GO Annotations Tab](#) tab.

Quantitative Analysis...

Scaffold includes a number of statistical tests that can be applied using various types of quantitative methods. These tests can be set up through the menu option **Experiment > Quantitative Analysis...**. When selected the dialog **Quantitative Analysis...** opens up showing the list of statistical tests available, normalization and quantitative methods options and two lists from which the User can choose the different categories he/she wants to compare and apply inference tests. There are up to seven tests potentially available, depending on the number of loaded samples and categories.

Figure 4-20: Quantitative Analysis Dialog Box



Other features:

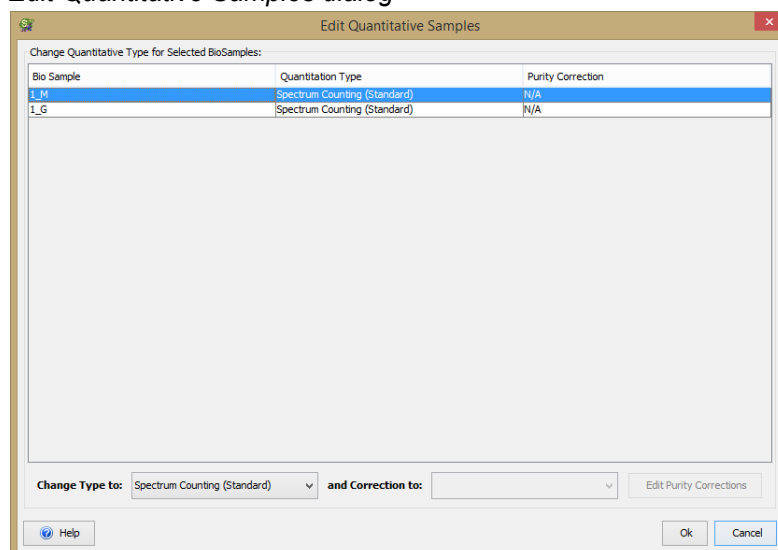
- **Use Normalization** check-box, see [Normalization among samples in Scaffold](#)
- **Minimum Value** pull down list
- **Quantitative Methods** pull down list, see [Label Free Quantitative Methods](#)

Note: When the **Use Normalization** check-box is not selected and Total Spectrum Count is the quantitative method chosen for the analysis, the values shown in the Samples View when the display option **Quantitative Value** is chosen are going to be the same as the one reported when **Total Spectrum count** is the selected Display Option.

•

Edit Quantitative Samples

The menu option **Quant > Edit Quantitative Method/Purity Correction** opens the **Edit Quantitative Samples** dialog. The dialog allows the User to change the quantitative methods selected for each BioSample when loading data into Scaffold

Figure 4-21: *Edit Quantitative Samples dialog*

The dialog includes a table listing all the different BioSamples included in the current Scaffold experiment, the quantitative method selected when loading and if relevant the related purity correction.

Below the table there are a couple of pull down lists and a button:

- **Change Type to:** - Pull down menu that lists the available selections for quantitation methods in Scaffold.
- **and Correction to:** - This pull down menu is available only when iTRAQ and TMT as selected as quantitative methods. It lists the Purity corrections tables available.
- **Edit Purity Correction** - This button is available only when iTRAQ and TMT as selected as quantitative methods. When selected it opens the [Edit iTRAQ/TMT Purity Corrections](#) dialog.

Changing the Quantitative type for a specific BioSample:

1. From the **Change type to** pull down list select a different quantitative method and then click **OK**.
2. When the quantitative type selected is either ITRAQ or TMT, the **and Correction to:** pull down list and the **Edit Purity Correction** button become available.
 - Select a correction from the list if available or select **Other...** to open the [Edit iTRAQ/TMT Purity Corrections](#) dialog from where the User can create new purity corrections tables or edit existing ones. This dialog can also be reached by clicking the **Edit Purity Correction** button.

Edit iTRAQ/TMT Purity Corrections

Every batch of iTRAQ or TMT reagents contains trace levels of isotopic impurities that need to be corrected. The correction factors, or purity values, are usually reported in the certificate

of analysis that comes with the iTRAQ or TMT reagents kit. They indicate the percentages of each reporter ion that have masses differing by -2, -1, +1 and +2 Da from the nominal reporter ion mass due to isotopic variants.

Note: It is strongly recommended to add these correction factors into Scaffold.

The **Edit iTRAQ/TMT Corrections** dialog opens when the **Edit Purity Corrections** button or the **Other...** option in the **and Correction to:** pull down list present in the [Edit Quantitative Samples](#) dialog are selected.

The dialog includes the Loaded Purity Corrections table which lists saved correction tables with their specific methods and a number of functional buttons appearing at the bottom of the table:

- **New Correction** - Opens the dialog [Purity Corrections](#) where the User can define a new purity correction table.
- **Edit...** - Opens the dialog [Purity Corrections](#) where the currently selected purity correction table is shown and where the User can adjust the values already included in the table or add others.
- **Delete** - Deletes the selected entries from the table
- **Close** - Closes the dialog without applying the changes
- **Apply** - Chooses the selected Purity Correction table.

Purity Corrections

The **Purity Correction** dialog opens when selected from the [Edit iTRAQ/TMT Purity Corrections](#) dialog through the buttons **New Correction** and **Edit...**

Figure 4-22: Purity Correction dialog

The dialog contains a matrix where the User can input or modify the isotope correction factors for iTRAQ or TNT. The percentages for each iTRAQ or TNT reagent need to be typed in following the same order as listed in the Certificate of Analysis. If the certificate of analysis is not available the User can use the Scaffold default values, although it is not recommended.

When a new Purity Correction table is created the User needs to assign a name to the table by typing one in the **Name** text box located above the matrix. Whether editing an existing Purity Corrections table or creating a new one, clicking **Apply** finalizes either one of the operations and closes the dialog.

Note: For more information about the way Scaffold calculates and applies iTRAQ corrections see the following publication: [Shadforth \(2005\)](#).

Referencing Scaffold

The User is free to copy, modify, and distribute the following examples for citing Scaffold in publications and reports.

Scaffold (Proteome Software Inc., Portland, OR 97219, USA) was used to probabilistically validate protein identifications derived from MS/MS sequencing results using the X!Tandem ([Craig \(2003\)](#)) and ProteinProphet computer algorithms ([Nesvizhskii \(2003\)](#)).

Scaffold (Proteome Software Inc., Portland, OR 97219, Oregon, USA) was used to validate protein identifications derived from MS/MS sequencing results. Scaffold verifies peptide identifications assigned by SEQUEST, Mascot or other search engines (list other search engines used to derive the imported data) using the X!Tandem database searching program ([Craig \(2003\)](#) and [Searle \(2008\)](#)). Scaffold then probabilistically validates these peptide identifications using PeptideProphet ([Keller \(2002\)](#)) and derives corresponding protein probabilities using ProteinProphet ([Nesvizhskii \(2003\)](#) and [Searle \(2010\)](#)).

IdentityE

Scaffold supports Waters' IdentityE (aka MS^E, aka hi-lo energy scanning). To be able to load data analyzed using PLGS into Scaffold, Proteome Software, in collaboration with Waters, developed a plug-in that comes with the PLGS installation. The plug-in specifically creates files compatible with Scaffold. Waters should have provided a Scaffold plug-in manual which guides the User through the Scaffold plug-in installation, but if this is not the case there is a copy available on Proteome Software website at [Scaffold4 PLGS plug-in](#)

The Scaffold plug-in only exports data that was searched in PLGS using MS^E.

Furthermore searches need to be run in PLGS with FDR set to 100% so that enough negative hits can be exported and available for Scaffold to be able to compute peptide and protein prophet probabilities in a statistically correct fashion.

When IdentityE data is loaded into Scaffold, an additional menu, IdentityE Menu, appears on the main menu bar after the Help menu. The menu provides options to configure absolute quantification.

Warning: Scaffold uses its own algorithms (Peptide/Protein Prophet, protein grouping) to determine both the list of proteins displayed in the Samples Table and their absolute quantities. While the intent is to reproduce Water's quantification strategy (top 3 peptides per protein), because of these algorithm differences, the list of proteins displayed and their quantities may differ somewhat from what's displayed in Protein Lynx. If you notice particularly large or confusing discrepancies, please do let us know.

Quantitation Option

Selecting the entry **IdentityE > Quantitation Option** opens the dialog **PLGS Quant Configuration**. The dialog contains:

- **Known Abundance Protein** pull down list - The list is used to select among the proteins listed in the Samples table the protein that has a known abundance input the value and use it for quantitative purposes.
- **Use accession not name** check box - Used to toggle the way the proteins are shown in the above pull down list.
- **How much?** text box - Used to input the quantitation normalization factor when the data is shown using weight or volume.
- **Select Unit for Showing Data** pull down menu - Used to choose the units of measurement for the quantitation. The Default value is intensity, calculated using Water's quantitation strategy (top 3 peptides per protein).





Tool-bar

Figure 4-23: Scaffold Tool Bar



The Scaffold tool-bar contains icons that represent equivalent commands for frequently used main menu options.

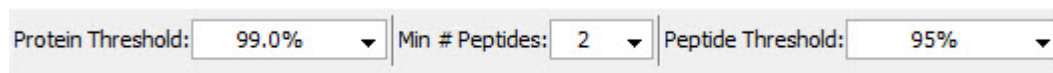
Icon	Function
	New —Initializes a Wizard which guides the User through the loading phase of the search data files in Scaffold. See The Loading Wizard
	Open —Opens a saved Scaffold experiment file, *.SF3, through a file browser.
	Save —Standard Windows behavior.
	Print —Prints the current view.
	Print Preview —Previews current view with the option to print the document.
	Copy —For each view copies to the clipboard the first table appearing at the top of the view. From there, the user can paste it into a third-party program such as Excel or Microsoft Word.
	Find —Opens a find dialog box that searches the first table present in the current view
	Excel —Exports the information that is contained in the current view to a tab-delimited text file that can be opened and viewed in Excel.
	BioSample Summarization level —See
	MS/MS Sample Summarization level —See.
	Add BioSample —Not available in the Viewer version, it initializes The Loading Wizard
	Queue Files for Loading —Not available in the Viewer version, see Queue Files for Loading .

Icon	Function
	Load and Analyze Queue —Not available in the Viewer version and active only when there are files listed in the loading Queue in the The Load Data View waiting to be loaded in Scaffold. When selected it opens the Load and Analyze Data page of the Loading Wizard.
	Quantitative Analysis —Opens the Quantitative Analysis... dialog
	Scaffold Q+/Scaffold Q+S —Available when running Scaffold Q+ or Scaffold Q+S, opens the Scaffold Multiplex Quantitation window.
	Help —Opens the Scaffold Online Help.

Filtering pane

The Scaffold Filtering pane, located on the right of the Tool-bar, contains filters and thresholding tools the User can adjust to increase or decrease the length of the displayed protein list in the Samples Table, see [Filtering Samples](#).

Figure 4-24: Scaffold Filtering Pane




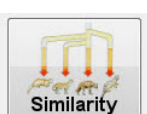


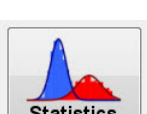


The image shows a horizontal toolbar with three dropdown menus. The first menu is labeled 'Protein Threshold:' and has a value of '99.0%'. The second menu is labeled 'Min # Peptides:' and has a value of '2'. The third menu is labeled 'Peptide Threshold:' and has a value of '95%'. Each menu has a small downward arrow on its right side.

Protein Threshold:	99.0%	Min # Peptides:	2	Peptide Threshold:	95%
--------------------	-------	-----------------	---	--------------------	-----

Navigation pane

The Scaffold Navigation pane is a vertical bar displayed on the left side of the Scaffold window. The bar contains buttons that toggle the seven different views available in the Scaffold main window.

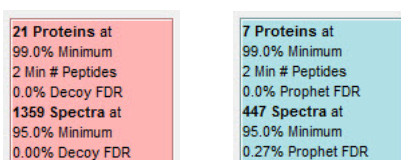
 <p>Load Data</p>	<ul style="list-style-type: none"> • See The Load Data View
 <p>Samples</p>	<ul style="list-style-type: none"> • See The Samples View
 <p>Proteins</p>	<ul style="list-style-type: none"> • See The Proteins View
 <p>Similarity</p>	<ul style="list-style-type: none"> • See The Similarity View
 <p>Quantify</p>	<ul style="list-style-type: none"> • See The Quantify View
 <p>Publish</p>	<ul style="list-style-type: none"> • See the Publish View
 <p>Statistics</p>	<ul style="list-style-type: none"> • See The Statistics View

FDR Dashboard

Scaffold calculates the False Discovery Rate (FDR) for both peptides and proteins and reports the values in the **FDR Dash Board** located underneath the navigation pane.

Protein and peptide FDR values are reported based on the specific protein and peptide thresholds selected in the [Filtering pane](#).

Figure 4-25: FDR Info Box - Red background: searches run with decoy concatenated database; Blue background: searches run with target database



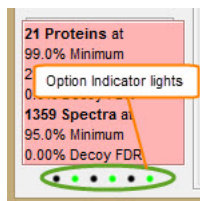
Depending on the type of database used to search the data loaded in Scaffold, the FDR is calculated in the following ways:

- When the search is performed against a target database, the FDR is calculated with proteins and peptides probabilities estimated using Peptide and ProteinProphet. The FDR dashboard where the values are reported appears with a blue background.
- When the search is performed against a decoy or reversed concatenated database, the FDR is calculated using the count of decoys against target identification hits. If proteins and/or peptides are filtered based on FDR, then the dashboard reports the specific protein and peptide thresholds necessary to reach those specific FDR values based on the FDR Browser landscape. The FDR box where the values are reported appears with a red background.

Option Indicator Lights

The Option Indicator Lights are six multi colored dots located at the bottom of the Navigation Pane underneath the [FDR Dashboard](#) in the Scaffold Main window.

Figure 4-26: Option Indicator Lights



Their scope is to remind the User about the status of the following options:

- View Menu Options* (green when selected):

- Show less <5% probability
- Show lower Scoring Matches
- Show entire protein Clusters
- *Load and Analyze Options* (green when selected):
 - Use Protein Cluster Analysis
 - Use Independent Sample Grouping strategy
- *Scoring Scheme*:
 - LFDR-green
 - PeptideProphet with Delta Mass correction - orange
 - PeptideProphet no mass correction -black

The User can always hover over each one of the colored dots to check their function; a tool tip appears providing a description of the selected dot.

Display pane

The information included in the different views appears in the Scaffold Display pane. Depending on the view, the type of information reported might appear framed in one or more tables or graphs included in one or more sub-panes. All panes and tables included in Scaffold share the following characteristics:

- Tool-tips
- Resizing of columns and panes
- Moving columns around
- Column sorting feature
- Multi selection of rows in the Samples Table

Tool-tips

The user can view information about fields or columns in a View by just hovering the mouse pointer over the location of interest. This operation opens a collapsed tool-tip. Pressing F2 opens an expanded tool-tip. Pressing the Escape (ESC) key on the keyboard closes the expanded tool-tip,

Figure 4-27: Viewing information in a collapsed tool-tip

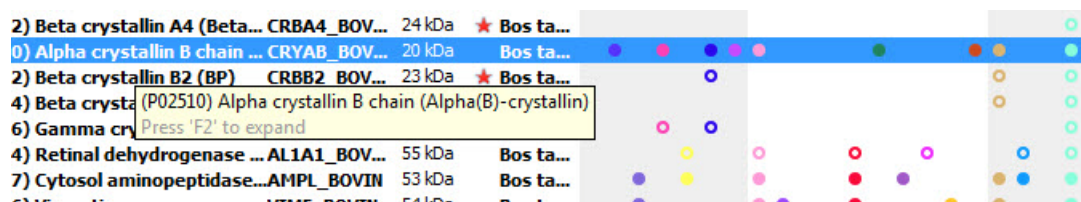
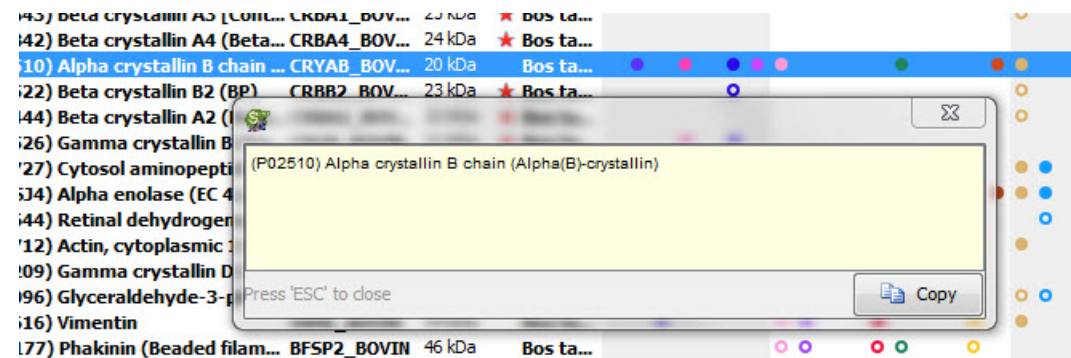


Figure 4-28: Viewing information in an expanded tool-tip

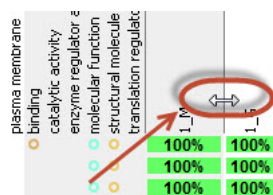


Resizing of columns and panes

The user can resize columns and different panes in each of the views to better suit his/her working needs. For example, in [The Samples Table](#), the user can change the width of a column by resting the mouse pointer on the right side of a column heading until the pointer

changes to a double-headed arrow, and then dragging the boundary until the column is the width that he or she wants.

Figure 4-29: Changing the width of a column in the Samples View



Moving columns around

In all tables throughout Scaffold, but the Samples Table, every column can be moved around from one position to another for more comfortable access to the data that is summarized in them.

The User simply has to click on the header of the column that he/she desire to move and drag it to the location where he/she wants to place it. Switching to another view will keep the columns in the new positions.

Figure 4-30: Moving columns around in tables

Valid	...	Sequence	Prob	SEQU	NTT	Modifications
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSSTPEQAAAGK(S)	100%	0.43	70	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSSTPEQAAAGK(S)	100%	0.43	77	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSSTPEQAAAGK(S)	100%	0.42	82	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSSTPEQAAAGK(S)	100%	0.47	00	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSSTPEQAAAGK(S)	100%	0.33	26	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(-)AEQHSSTPEQAAAGKSHGGLGGSYK	100%	0.35	37	2 Acetyl (+42)
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.41	11	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.32	28	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.35	64	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.26	31	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.41	10	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.37	30	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.36	04	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.40	26	2
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	(K)SHGGLGGSYK(V)	100%	0.41	11	2

Column sorting feature

In all tables throughout Scaffold, the User can use the tri-state column sorting feature and sort the display by clicking on any column header. For example, to sort the proteins based on increasing molecular weight, the User can click the Molecular Weight column header once. To sort the proteins based on decreasing molecular weight, the User can click the Molecular Weight column header twice. To return to the default display, the User can click the Molecular Weight column header a third time.

Multi selection of rows in the Samples Table

In the Samples table the User can select multiple rows by using either the SHIFT or the CTRL key, depending whether the desired selection has contiguous rows or not, and the click of the mouse in a pretty standard fashion. Other functions can then be applied, like assigning a star to the selected group of proteins in the Samples table, for example.

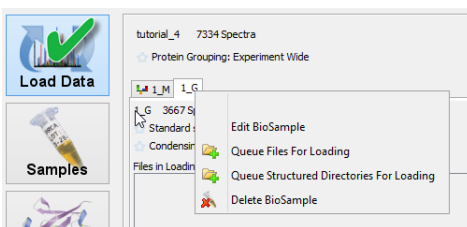
Mouse Right Click Menus

When the User right clicks the mouse while hovering over the Display Pane, a menu with various options appears close to the working arrow. Depending on the current view the list of options available in the menu varies. A description of the mouse right click command is provided in [“Mouse Right Click Commands” on page 212](#).

Load Data View

The following menu appears when the User right clicks on the BioSample name tab.

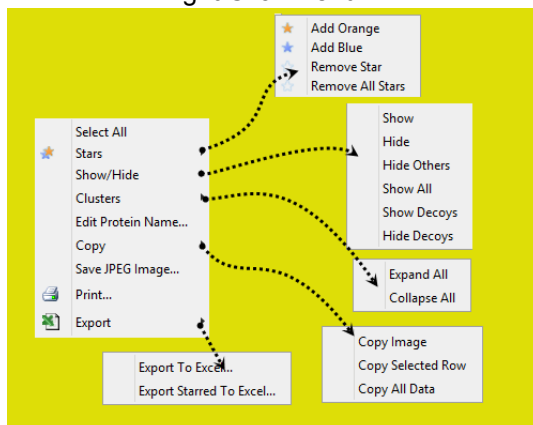
Right Click Menu A:



Samples View

When the User right clicks anywhere over the list of proteins the following menu appears which contains a number of sub-menus:

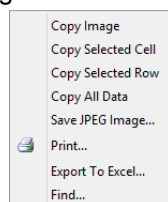
Right Click Menu B:



Proteins View

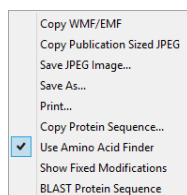
When the User right clicks over the Proteins View generally [Right Click Menu C](#) appears, but depending on the selected tab a different menu might be available.

Right Click Menu C:



- Protein Sequence tab - Available right click menu:

Right Click Menu D:

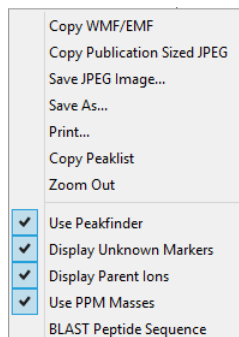


Chapter 4

The Scaffold Window

- Spectrum Tab and Spectrum/Model Error - Available right click menu:

Right Click Menu E:



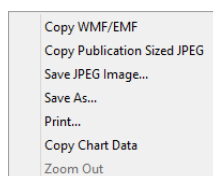
Similarity View

- Grouping Table - Available right click menu: [Right Click Menu C](#)
- Identifications Tab and Fragmentation Table tab - Available right click menu: [Right Click Menu C](#)
- Spectrum and Spectrum/Model Error tab - Available right click menu: [Right Click Menu E](#)

Quantify View

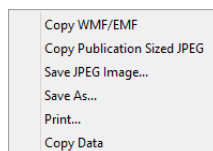
- Quantitative Value Pane and Quantitative Scatter-plots - Pane Available right click menu:

Right Click Menu F:



- Protein Venn Diagram Pane - Available right click menu:

Right Click Menu G:



When clicking on a Venn diagram set a list of the proteins or unique peptides or spectra will appear. When mousing on the list and right clicking on it [Right Click Menu C](#) appears.

- Gene Ontology Terms Pane

When mousing over in any of the tabs available in this pane [Right Click Menu C](#) appears.

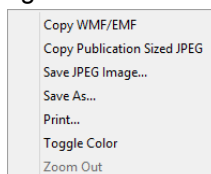
Publish View

When mousing over the Experiment Methods tab [Right Click Menu C](#) appears.

Statistics View

When mousing over the MS/MS samples table [Right Click Menu C](#) appears. When mousing over the various graphs found in the different panes [Right Click Menu F](#) appears. When mousing over the FDR browser the available right click menu is:

Right Click Menu H:



Chapter 4

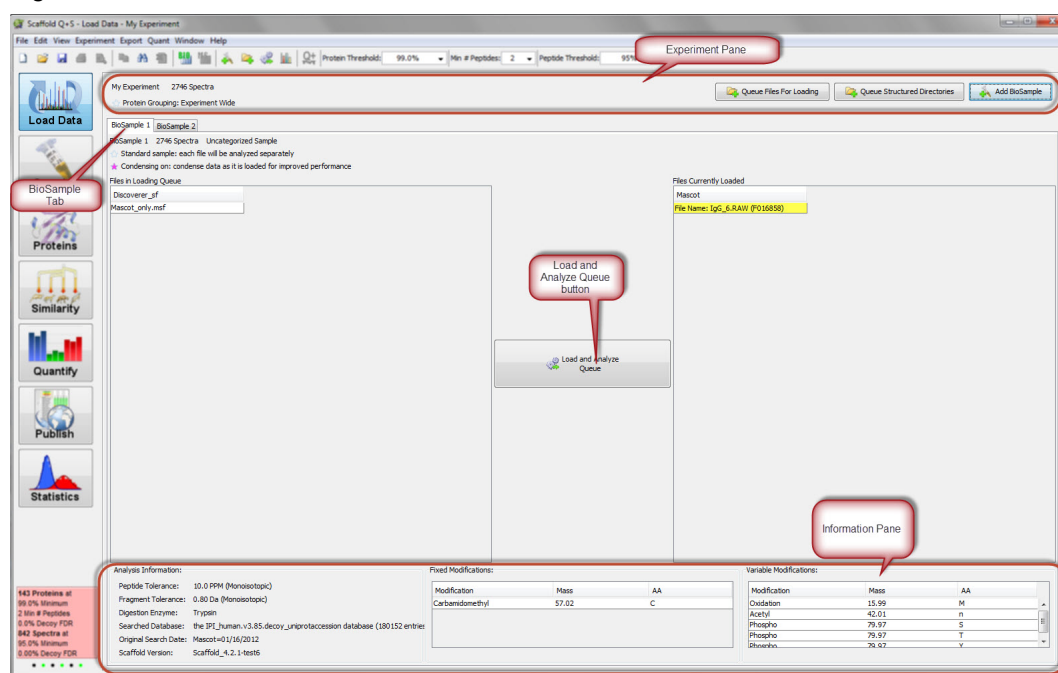
The Scaffold Window

Chapter 5

The Load Data View

Scaffold's Load Data View provides an overview of the currently opened experiment together with tools for loading further MS files or deleting them or adding or deleting BioSamples. Through this view the User can see and check the list of files loaded in each BioSample; add or delete BioSamples and MS samples; check their analysis information, the fixed and variable modifications or edit BioSample information.

Figure 5-1: The Load Data View



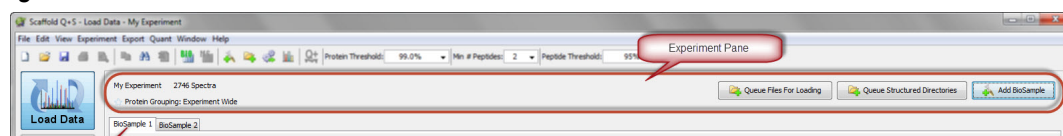
This chapter details the Scaffold Load Data View describing the different elements that constitute it:

- “[Experiment Pane](#)” on [page 112](#), which provides general information about the currently loaded experiment and tools to add MS samples to a BioSample or create a new BioSample within the current experiment.
- “[BioSample tabs](#)” on [page 117](#), which contain the lists of already loaded MS files or files in queue for each BioSample together with specific loading information.
- “[Information pane](#)” on [page 119](#), which provides specific in depth information about the search files loaded in a specific BioSample.

Experiment Pane

The Experiment Pane provides general information about the currently loaded Scaffold experiment.

Figure 5-2:



On the left side of the pane Scaffold shows the name of the current experiment, which by default is called My Experiment, the total number of spectra currently loaded in the experiment and the type of grouping selected at the time of loading.

There are two types of grouping modes implemented in Scaffold:

- **Experiment wide** - Scaffold groups proteins across all MS samples and BioSamples.
- **Independent sample** - Scaffold groups proteins only within each MS sample. Each MS sample appears as if it was loaded independently.

In the top right portion of the pane, not appearing in the Viewer mode, there are three buttons:

- [Queue Files for Loading](#) - Adds more MS samples to the selected BioSample
- [Queue Structured Directory for Loading](#) - Adds more MS samples organized in separate directories.
- **Add Biological Sample** - Adds a new BioSample to the experiment by starting the [The Loading Wizard](#)



*Files can also be added to a BioSample directly from the Mascot Server. The User can do so by selecting a BioSample and going to the menu option **Experiment** > **Queue Files from Mascot Server**, see [Queue Files From Mascot Server for Loading...](#)*

Queue Files for Loading

Selecting this command opens a standard file browser. From there the User can navigate to the location where the data files to be loaded in Scaffold are stored. Once the files are selected Scaffold places them in the Loading Queue

The **Queue Files For Loading** command can be selected from the following locations in the program:

- The Experiment menu
- The Load data View

- The [Queue files for loading](#) page in the Wizard

The User should be able to reach the files of interest from the system where Scaffold is installed and should also make sure that the format of the data files is supported by Scaffold by consulting the [file_compatibility_matrix.pdf](#).

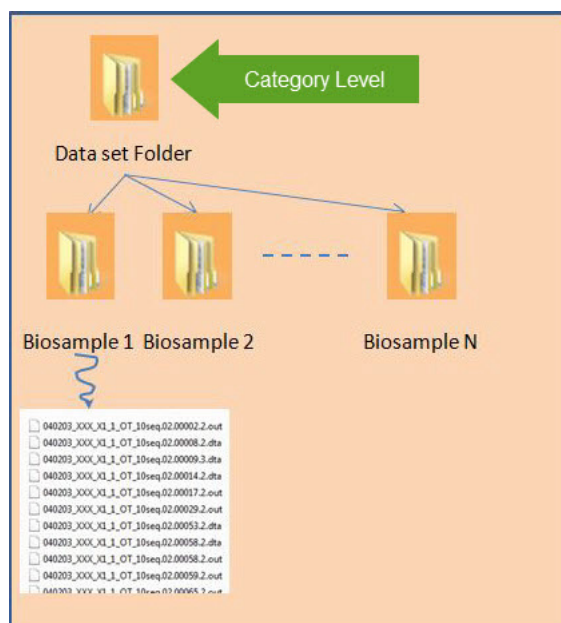
When selected outside of the Loading Wizard and if the User has multiple biological samples already defined, he/she should make sure that the right biological sample is chosen in the Load Data view before beginning to queue files.

Multiple files can be selected at one time as long as they contain search results of data run against the same FASTA database. The User is asked to specify the database in the [Load and Analyze Data](#) page of the Wizard.

Queue Structured Directory for Loading

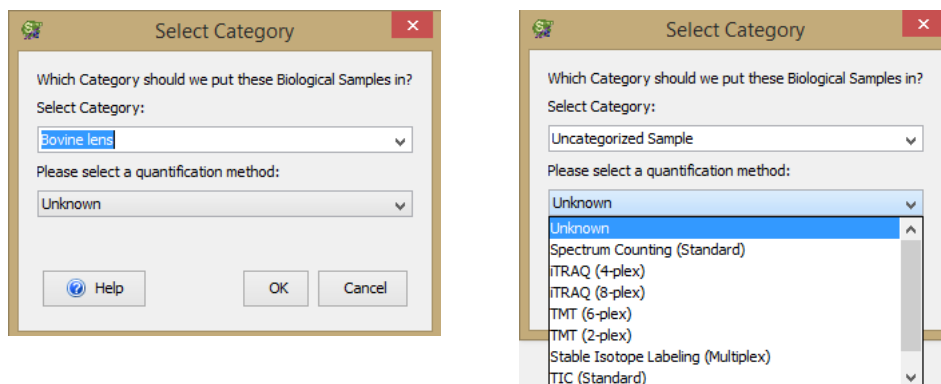
This command allows the User to streamline the loading of groups of data files organized in a number of directories and sub directories for a specific category. The organization of the directories should be similar to the one depicted in [Figure 5-3](#).

Figure 5-3: Organization of structured directories



Clicking the button **Queue Structured Directories** opens a file browser. The User should then point Scaffold to the top level of the structured directory containing the data files to be loaded in a specific Category. Scaffold loads all the MS files found in one of the second level folders in a single BioSample, even when the files are contained in sub-sub-folders.

Once the top level folder is selected a dialog opens asking to define the name of the category. If the User is running Scaffold Q+ or Scaffold Q+S the dialog also asks to define the type of quantitation to perform with the data.

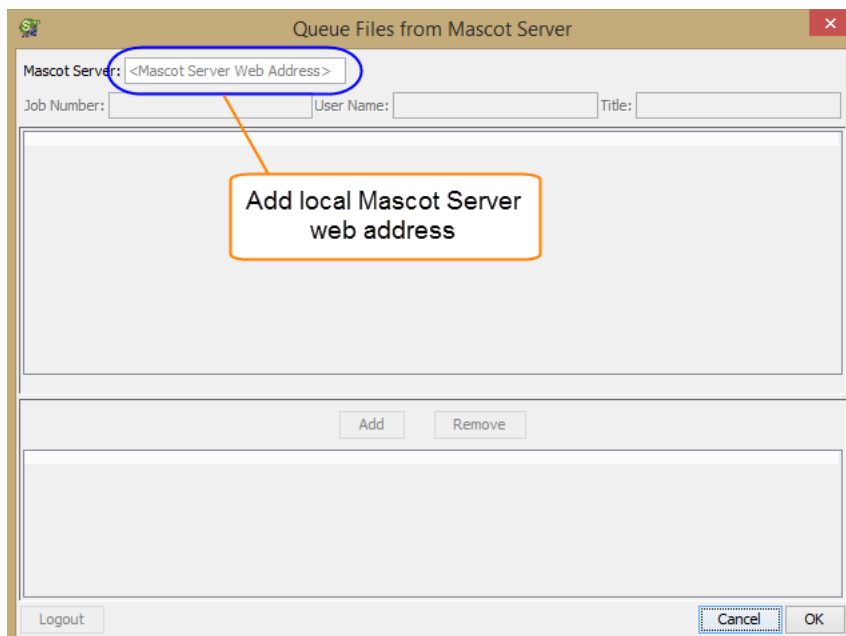
Figure 5-4: Defining Categories and quantitation methods

If the data is already organized in different categories folders, the User can load one category at the time by pointing Scaffold to one specific category folder. All the different sub-folders included in the selected folder are assigned to biosamples that will all appear under that particular category.

Queue Files From Mascot Server for Loading...

The User can open this dialog either from the **Scaffold Wizard** on the Queue Files for Loading page or from the menu item **Experiment > Queue Files From Mascot Server for Loading**. The dialog contains tools to connect Scaffold to a Mascot Server, select and download searched data files directly from there into Scaffold.

When calling the dialog from the menu item **Experiment > Queue Files From Mascot Server for Loading** the User should make sure to have chosen the appropriate Bio/MS Sample before adding data.

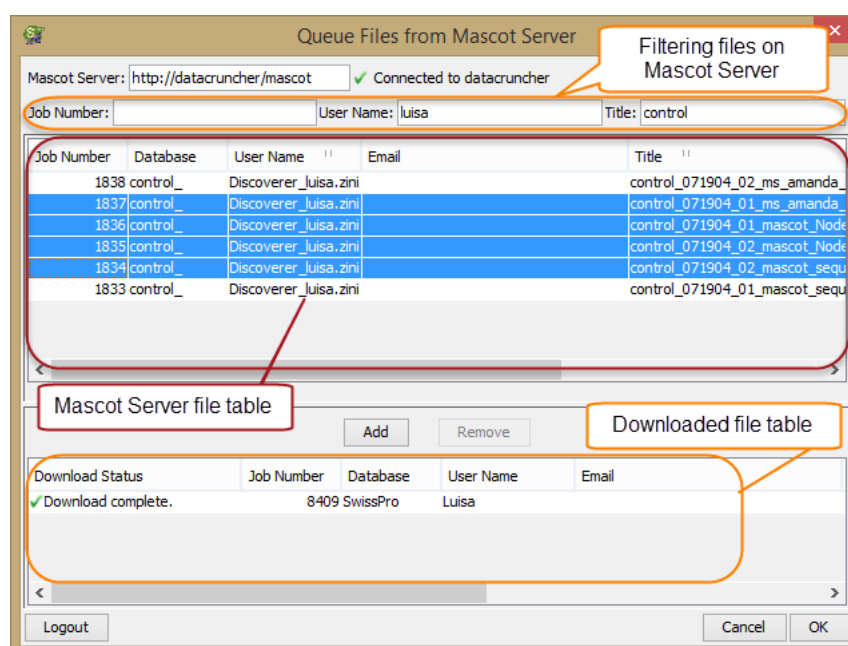
Figure 5-5: Connecting to the Mascot Server from Scaffold

When the User opens this dialog for the first time, he/she needs to connect to the local Mascot server. This is done by adding the local Mascot Server web address in the Mascot Server text box located in the top left corner of the dialog. If no security is implemented, Scaffold connects directly to the Mascot Server showing a list of files available for download.

If security is enabled on the Mascot server, a login window pops up asking for account name and password. The User should make sure that the account he/she is using in Mascot has administrative privileges.

Tip: **Edit > Preferences > Mascot Server** allows the user to create a default connection to a Mascot Server of choice. Scaffold automatically logs in to the server specified in the settings.

Figure 5-6: Queue files from Mascot Server for Loading



The dialog can be divided into 5 different panes containing a number of tools to help the User smoothly select and load data files into Scaffold.

- **Connection and filtering pane** - which contains information about the status of the connection to the Mascot Server spelled out in the **Mascot Server:** address text box. When the server is not connected this is the only text box available in the dialog, see [Figure 5-5](#). Below the connection information there are three different filters that can be applied to the data files shown in the Mascot Server file table. This helps the User quickly locate the files he/she wants to load into Scaffold. The available filter are:
 - *Job number*
 - *User name*
 - *Title*
- **Mascot Server file table**- When connected, the table lists the search data files saved on the server. The table shows the typical functionalities described in the [Display pane](#)

section plus it accepts bulk operations like the standard windows multiple selection of files. The filtering pane acts on this table so that the User can easily locate the files he/she wants to load in Scaffold.

- **Action pane** - Contains two buttons. The **Add** button: active only when there are files selected from the **Mascot file table**. It starts the download of the chosen data files from the Mascot server to the computer where Scaffold is running and adds them to the loading queue in the Load Data View. The **Delete** button: active only when there are files selected in the Download file table. When clicked it deletes the highlighted files.
- **Downloaded file table** - The table lists the files downloaded from the Mascot Server to the computer where Scaffold is running. The table shows the typical functionalities described in the [Display pane](#) section plus it accepts bulk operations like the standard windows multiple selection of files. The status of the download is reported in the Download status column when completed a green check appears.
- **Completion pane** - Contains three buttons. The **Logout** button: the User can use this button to logout of the current Mascot Server and login to another one. The **Cancel** button: standard Windows functionality. The **OK** button: to finalize the loading of the downloaded files.

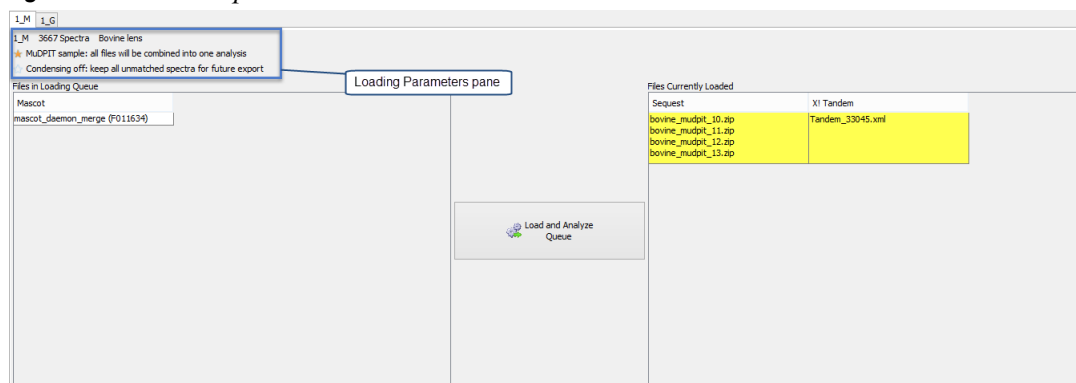


*If by any chance the User is logged into the Mascot Server as a Guest, Scaffold is does not accept to download files and shows an error. When this happens, to be able to access the Mascot login window again, the User has to clear the address in the **Mascot Server:** address text box and then press enter. The Mascot Login dialog opens allowing the User to login with a different account.*

BioSample tabs

In the Samples View each BioSample, defined in the currently opened experiment, has a specific tab window assigned to it. The tab window is labeled with the same name as the BioSample and contains information about the loading status of the experiment, the MS experiments loaded or about to be loaded into the BioSample and which option they were chosen at the time of the load.

Figure 5-7: BioSample tabs in the Load Data View



The elements included in the BioSample window are the following:

- **Loading Parameters pane** - The information included in this pane reports the name of the BioSample, the number of spectra loaded and the name of the category. Underneath it reports the settings selected during the loading phase on the New BioSample page of the Wizard. This means whether standard (separate processing for each MS sample) or MuDPIT (combined analysis for all samples) applies and whether the samples were loaded using the condensing option or not. The User can change these settings by editing the BioSample: from the Experiment menu or by right clicking the BioSample tab, see [Edit BioSample](#).
- **Files in Loading Queue table** - Lists the files ready to be loaded in Scaffold. If the User has selected files from more than two search engines, he/she needs to scroll towards the right side of the table to see all the files.
- **Files Currently Loaded** - Lists the files already loaded in Scaffold. The loaded files are highlighted in yellow when the Scaffold analysis is completed. After analysis, files from the same MS sample run through multiple search engines, are aligned on a single row. Hovering the cursor over a file shows its full name (see [“Mascot File Names”](#) on page 209) and the database loaded.
- **Load and Analyze Queue button** - Opens the Load and Analyze Data page of the Loading Wizard, see [Load and Analyze Data](#).

The User can modify the name of the BioSample and categories by opening the dialog [Edit BioSample](#). This dialog can be reached through the Experiment menu or by right-clicking the mouse over the BioSample tab, see [Samples View](#) in the [Mouse Right Click Menus](#)

Chapter 5

The Load Data View

section.

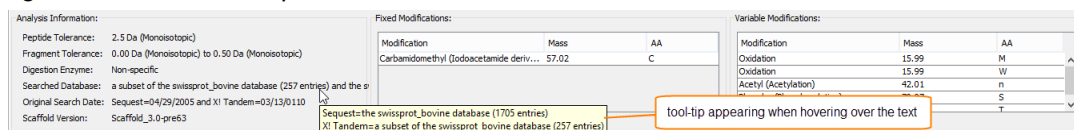
Note: The User can delete files before analysis, or MS samples after analysis, by right-clicking and selecting Remove Selected Samples.

Information pane

The bottom section of the Load Data view contains three information panes. Each of them provides information related to how the loaded data was analyzed by the search engine:

- [The Analysis Information pane](#)
- [The Fixed Modifications pane](#)
- [The Variable Modifications pane](#)

Figure 5-8: Information panes in the Load Data view



The pieces of information provided in the panes describe the data contained in the files listed in the table **Files Currently Loaded**. If specific files are highlighted in this table, then the information is restricted to the highlighted files. Otherwise it describes all the files in the displayed BioSample.

In contrast to these information panes which describe only one sample at the time, the Publish view summaries this analysis information for all the samples.

The Analysis Information pane

The Analysis Information pane lists the peptide and fragment mass tolerances, the digestion enzyme and the database searched and when that search was done. The Scaffold version is the version in place when the data was loaded into Scaffold. If several files are selected and these files have different parameters, this box shows the range of values. When holding the cursor over the data, a tool-tip shows further details.

The Fixed Modifications pane

This pane contains a table listing the fixed modifications, with their masses and the related modified amino acids, used during the searches recorded in the loaded files belonging to a specific BioSample.

The Variable Modifications pane

This pane contains a table listing the variable modifications, with their masses and the related modified amino acids, used during the searches recorded in the loaded files belonging to a specific BioSample.

Note: When a peptide starts with E or Q, X!Tandem automatically checks for the formation of pyroglutamic acid, i.e., the loss of water or ammonia, respectively. The Pyro-Glu modification then automatically appears in the table, see [Analyze with X!Tandem Pane](#).



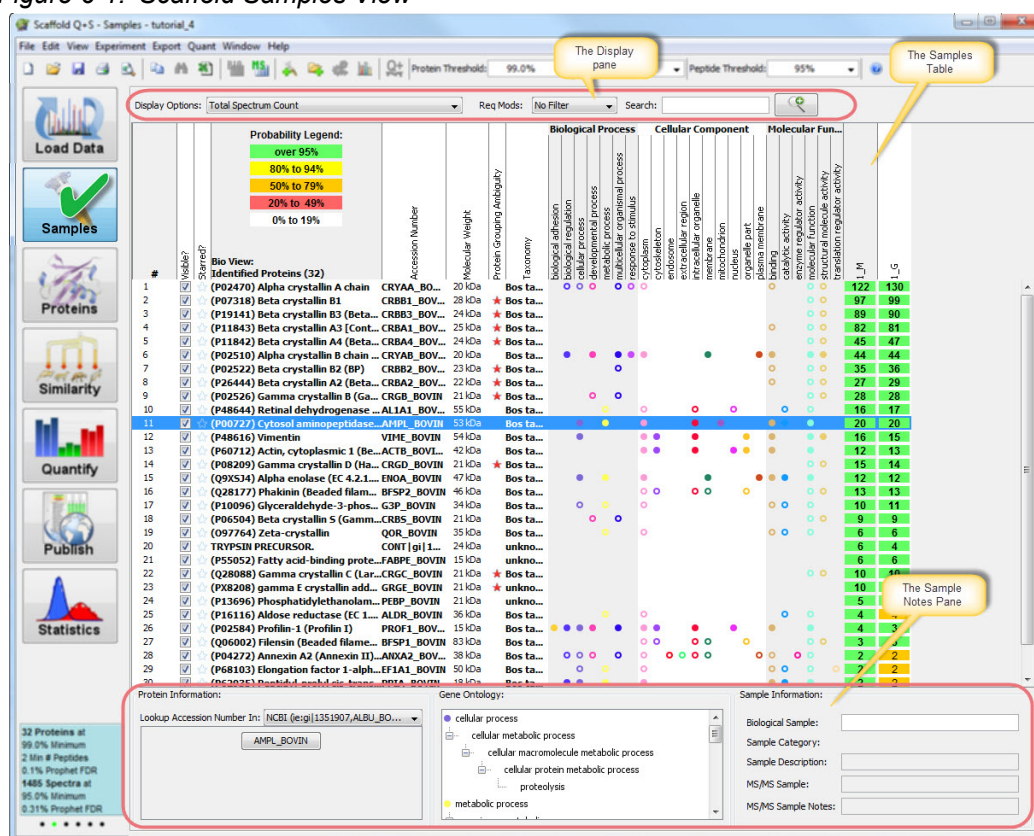
*The fixed and variable modifications are those used in the searches for the files listed in the **Files Currently Loaded** list. If different modifications were used to create different files, the modifications that are not true for all files are highlighted in red. The User can see which files a red modification applies to by hovering over it.*

Chapter 6

The Samples View

The Scaffold's Samples View provides overviews and tools to help the User make direct comparisons among BioSamples or MS Samples regarding the content of identified proteins.

Figure 6-1: Scaffold Samples View



This chapter details the Scaffold's Samples View providing a description of the different elements that constitute it:

- [“The Samples Table” on page 123](#), which displays a summary of the experiment's results. The list of identified proteins appears as rows, and the list of BioSamples or MS Samples appears as columns.
- [“Filtering Samples” on page 129](#), which describes how to increase or decreases the number of proteins listed in the Samples Table.

Chapter 6

The Samples View

- [“The Display pane” on page 135](#), which provides options to view rough estimates of differential expressions. Scaffold uses a multitude of statistics to filter required modifications, providing both simple and advanced search options.
- [“Information Panes”](#), which lay out and specify useful protein, Gene Ontology, and sample information.

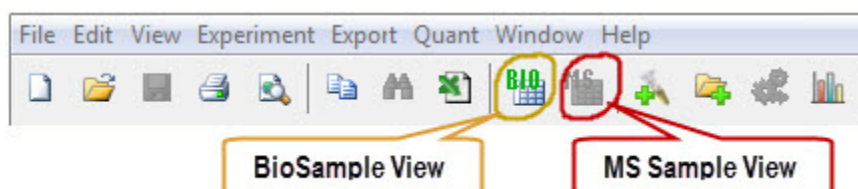
The Samples Table

When the Samples View first opens, all proteins that meet the default threshold settings are listed in the Samples Table. There are two levels of summarization the User can use to view the Samples Table offering two different ways of looking at the results:

- “The BioSample View”, which provides a single column overview of all the proteins groups or clusters in a given BioSample.
- “The MS/MS Sample View”, which displays protein identifications in separate columns by mass spectrometry sample

The two summarization views can be toggled using the BIO and MS buttons located underneath the main menu bar.

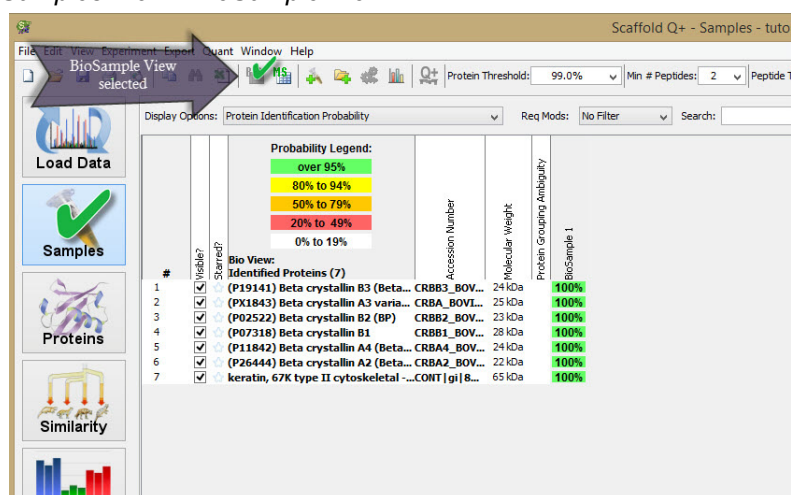
Figure 6-2: Scaffold Samples View - BioSample View/MS Sample View toggle buttons



The BioSample View

This view combines all MS Samples into a summarized BioSample level, which is the highest overview level of the results. Each BioSample is represented by a column, sorted by category and then by BioSample name.

Figure 6-3: Samples View - BioSample View

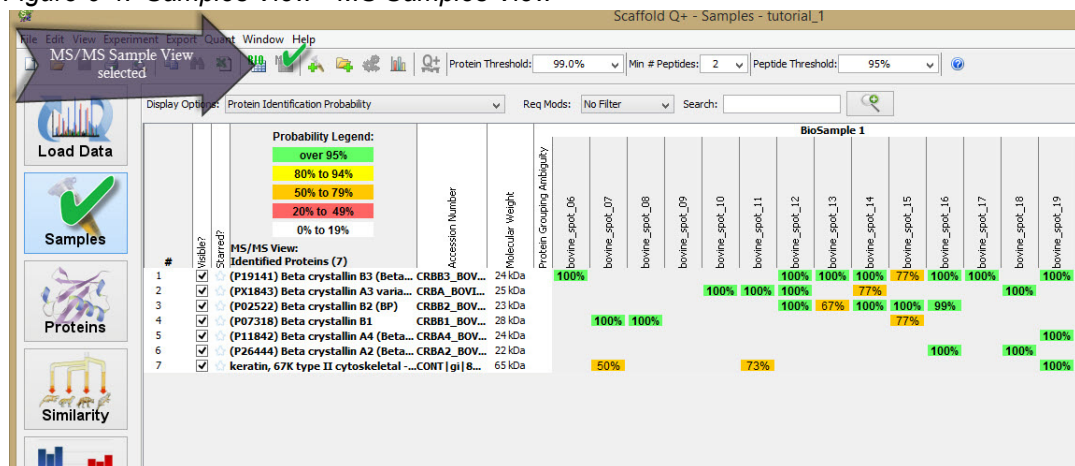


The MS/MS Sample View

When this view is selected, the Samples Table shows one column for each MS Sample, sorted first by category and then by BioSample. It is useful, for example, for analyzing samples processed with gels.

Scaffold displays a row in the table for each protein group or protein cluster which has at least one MS Sample identified that passes the assigned filter thresholds requirements. To display MS Samples that do not meet the confidence requirements, the User can select [Show Lower Scoring Matches](#) from the View Menu.

Figure 6-4: Samples View - MS Samples View



MS Sample vs BioSample summarization levels

Data associated with a BioSample might come from a sample taken by a doctor, medical researcher, or biologist, such as a drop of blood or tissue from an organism.

Using such techniques as 2D gels or liquid chromatography, proteins or peptides from these BioSamples are then separated from each other. Each resulting individual band, spot, or LC fraction then processed by a mass spectrometer is one mass spectrometry sample (abbreviated as MS sample).

One BioSample is therefore typically made up of more than one MS sample - sometimes many more.

Protein List

To simplify the inspection of the identified proteins, Scaffold aggregates the protein list using two levels of hierarchy.

- **Protein Group** - a group of proteins with identical sets of peptides.

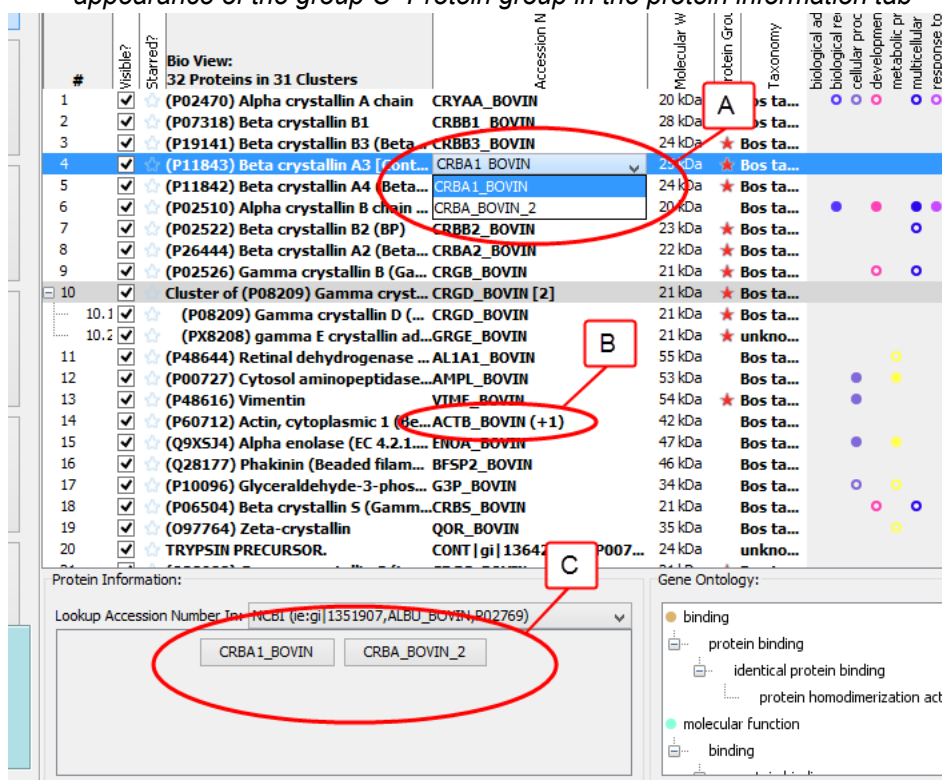
In the Protein list the protein groups are displayed collapsed, the number of proteins in the group is indicated in parenthesis close to the accession number of the protein

representing the group, see B in Figure 6-5. By default, the protein that has the highest probability and the most associated number of spectra will represent the group in the list.

When clicking on the accession number for a protein group a pull down list becomes available, the User can thus selected a different protein to represent the group in the protein list see A in Figure 6-5. It is also possible to over all change the proteins representing the groups by going to [Apply Protein annotation Preferences](#).

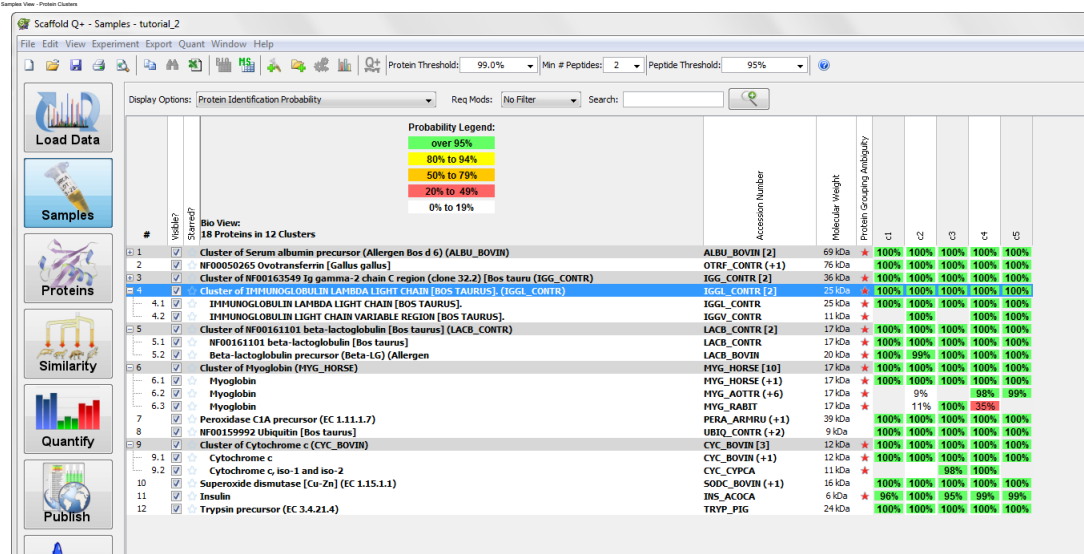
The different proteins present in the group are also listed in the [Protein Information pane](#) represented as buttons labeled with each protein accession number. By clicking one of the buttons is possible to gather further information looking up the proteins on specific look up sites.

Figure 6-5: Samples View - Protein groups -A Pull down list of proteins in the group. B regular appearance of the group C- Protein group in the protein information tab



- Protein Cluster** - a set of protein groups created using a hierarchical clustering algorithm. The clustering algorithm is similar to the one used by Mascot to create protein families, but with more stringent grouping rules. Members of the cluster share some peptides but not all of them. Protein clusters are by default represented by the protein group that shows the highest associated probability. Clusters can be collapsed or expanded directly in the protein list, see Figure 6-6. For more information about Scaffold clusters see Chapter 8, “Protein Grouping and Clustering,” on page 151

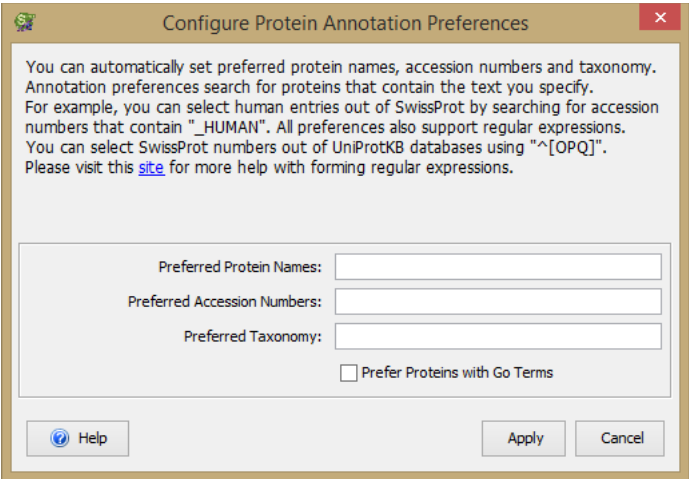
Figure 6-6: Samples View - Protein clusters



Apply Protein annotation Preferences

The menu option **Experiment > Apply Protein annotation Preferences** opens the dialog **Configure Protein Annotation Preferences** where the User can globally define which protein in a protein group is visible in the protein list appearing in the Samples Table.

Figure 6-7: Configure Annotation Preferences Dialog



The dialog provides a series of text boxes where the User can input his/her preferences.

By default Scaffold automatically selects the visible protein relying on the following five criteria in the order shown below:

1. Prefer proteins that contain sequences (user cannot modify this preference)
2. Prefer the accession number preference

3. Prefer the protein name preference
4. Prefer the taxonomy preference
5. Prefer proteins that contain GO terms

Probability Legend

To provide a measure of how correct protein identifications are for any of the BioSamples or MS Samples, Scaffold uses a couple of different validation algorithms which assign identification probabilities to the peptides. After that, using ProteinProphet, it groups the peptides by their corresponding protein(s) to compute probabilities that those proteins were present in the original sample, see [ProteinProphet](#). When loading data through the Wizard the User is presented with the following peptide validation scoring systems to choose from:


- **PeptideProphet Scoring** - This scoring algorithm learns the distributions of search scores and peptide properties among correct and incorrect peptides and uses those distributions to compute for each peptide a probability that it is correct, see [PeptideProphet](#).
- **LFDR-based Scoring** - This is a novel scoring algorithm based on a Bayesian approach to local False Discovery Rate. It is especially effective for QExactive and other high mass accuracy data., see [LFDR-based scoring system](#).

The protein probability values are reported in the Samples Table when selected from the Display Options pull down list. They are color coded to highlight significant differences in protein identification confidence. The coloring is kept even when another statistics is selected from the Display Options list. Located at the top of the view, the Probability Legend defines the color coding for the protein identification probability.

Sorting feature

When the Samples View first opens, the displayed proteins are initially sorted based on a protein probability of 50% (Scaffold's calculated probability, which is a percentage, that the protein identification is correct), and if any proteins have the same probability, then the proteins are sorted alphabetically based on their accession numbers. You can use the tri-state column sorting feature and sort the display by clicking on any column header. For example, to sort the proteins based on increasing molecular weight, click the Molecular Weight column header once. To sort the proteins based on decreasing molecular weight, click the Molecular Weight column header twice. To return to the default display, click the Molecular Weight column header a third time.

Proteins of Interest

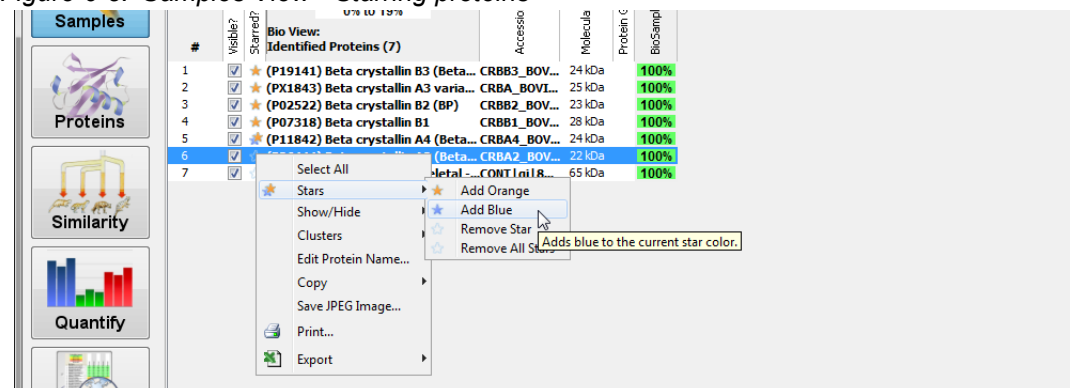
The User can mark proteins in an experiment that are of special interest by clicking the Star icon  in the Starred? column for the protein. Two different colored stars, blue and orange, and a combination of them are available by clicking multiple times on the same star

Chapter 6

The Samples View

or by selecting in the right click menu the option star. By using a combination of different stars it is possible to create three different sets of proteins of interest. You can then bring these proteins to the top of the display by clicking the Starred? column header. To return to the default protein display, click the column header twice more.

Figure 6-8: Samples View - Starring proteins



Hidden Proteins

If proteins that are not of any interest to the User are displayed in the Samples View, and/or contaminants are displayed, the User can remove these proteins from the view. To hide the entire protein entry in the Samples View, the User can simply clear the Visible option for the protein. For example, to eliminate Trypsin products from the view, the User can carry out a search for all proteins that contain “Trypsin” in their names, and then clear Visible option for all the proteins that meet this search criteria. Only those proteins that do not have “Trypsin” in their names are displayed.



To display the proteins that are hidden go to the menu View and toggle the menu entry Show Hidden Proteins

Protein Grouping Ambiguity

In the Samples View, a star in the column Protein Grouping Ambiguity indicates that the protein in this row is associated with one or more other proteins that share some, but not all, of their peptides. This visual clue marks the proteins for which it may be worthwhile to examine the shared proteins in the Similarity View. The stars in the Protein Grouping Ambiguity column are red when Scaffold loads the data. The stars turn green as a reminder that the User has already examined the Similarity view for the protein.

Filtering Samples

There are three different filters that can be used to increase or decrease the length of the displayed protein list in the Samples Table. Their function is to set minimum characteristics for identification confidence:

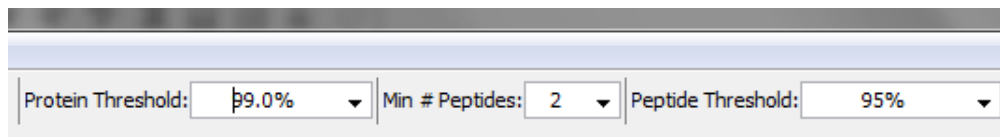
Protein Threshold

Minimum Number of Peptides

Peptide Thresholds

The protein and peptide thresholds filter probabilities or FDR values if the loaded data were searched using decoys. The drop down lists includes the two options depending on the type of searches loaded into Scaffold. It is possible to type a custom FDR threshold directly into the box by adding “FDR” to the end of the string, e.g. “10.3% FDR”, for more information see [FDR Filtering](#).

Figure 6-9: Scaffold Confidence Filters

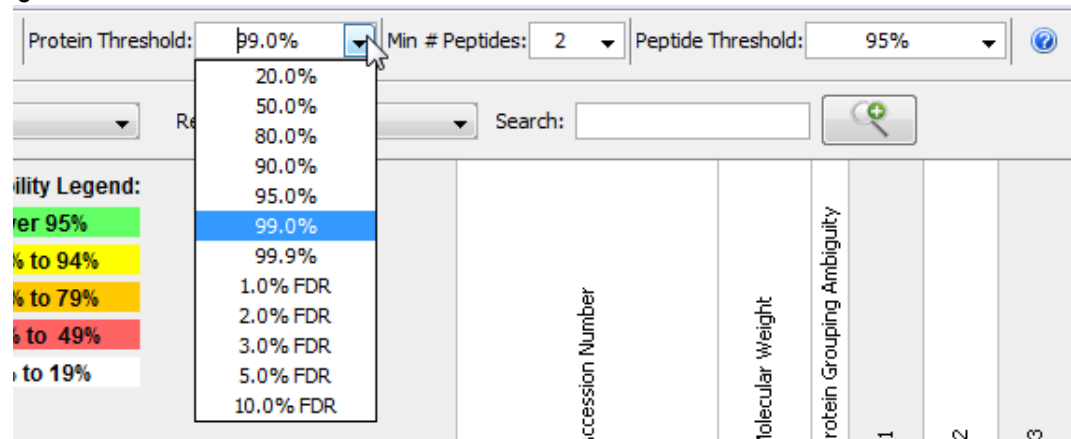


- Proteins are displayed if each of the filter options is met by at least one sample.
- Filters can be locked with a password. When locked, the filters cannot be changed unless the password is entered. This allows you to control what proteins are displayed when you distribute a Scaffold file.
- Also note that Protein probability is derived in part from peptide probability, so setting the protein probability much lower than the peptide probability likely won't display any more results

Protein Threshold

Through this pull down list the User can set the minimum requirement for Scaffold's calculated probability of correct protein identification. When the data loaded in Scaffold has been searched against a decoy database, [FDR Filtering](#) options become available as well, see [Figure 6-10](#).

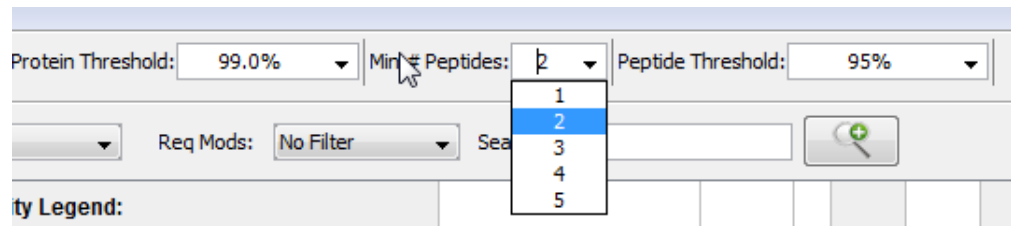
Figure 6-10: Protein Threshold



Minimum Number of Peptides

Through this pull down list the User can set the number of unique peptides that must be found for one protein in order to consider the protein to be identified.

Figure 6-11: Minimum Number of Peptides



Peptide Thresholds

Through this pull down list the User can set how certain a peptide identification must be before it can be counted toward the minimum number of peptides. When the data loaded in Scaffold has been searched against a decoy database, [FDR Filtering](#) options become available see [Figure 6-12](#).



This filter setting affects not only which proteins are shown but also the reported values shown for number of Exclusive Unique Peptides, number of Total Unique Spectra, Number of Exclusive Unique Spectra, and Percent of Total Spectra.

Among the entries for this filter, shown in the drop down list, the selection **Custom...** allows defining peptide filters based on the underlying search engines scores. See [Custom Peptide Filters](#) for more information regarding this option.

Figure 6-12: Peptide Thresholds

Protein Threshold: 99.0% | Min # Peptides: 2 | Peptide Threshold: 95%

Req Mods: No Filter | Search:

ity Legend:

- 95%
- to 94%
- to 79%
- to 49%
- to 19%

Accession Number

Molecular Weight

Peptide Threshold options:

- 0%
- 20%
- 50%
- 80%
- 90%
- 95%
- Custom...
- 0.1% FDR
- 0.5% FDR
- 1.0% FDR
- 2.0% FDR
- 5.0% FDR

Custom Peptide Filters

The option Custom peptide filters provides a way the create peptide filters based on the underlying search engines scores. When the User chooses custom filters, Scaffold ignores the protein probability and filters the proteins exclusively on the number of peptides that pass the selected custom peptide filter.

Custom filters can be created by selecting “Custom...” from the Peptide Threshold drop down list, see [Figure 6-12](#), or by going to the menu option **Edi t> Edit Peptide Thresholds** and open the Edit Peptide Threshold dialog, which shows a list of existing custom filters. The dialog allows either to edit an existing threshold, create a new set of parameters, or delete selected entries, see [“Configure Peptide Thresholds Dialog Box”](#).

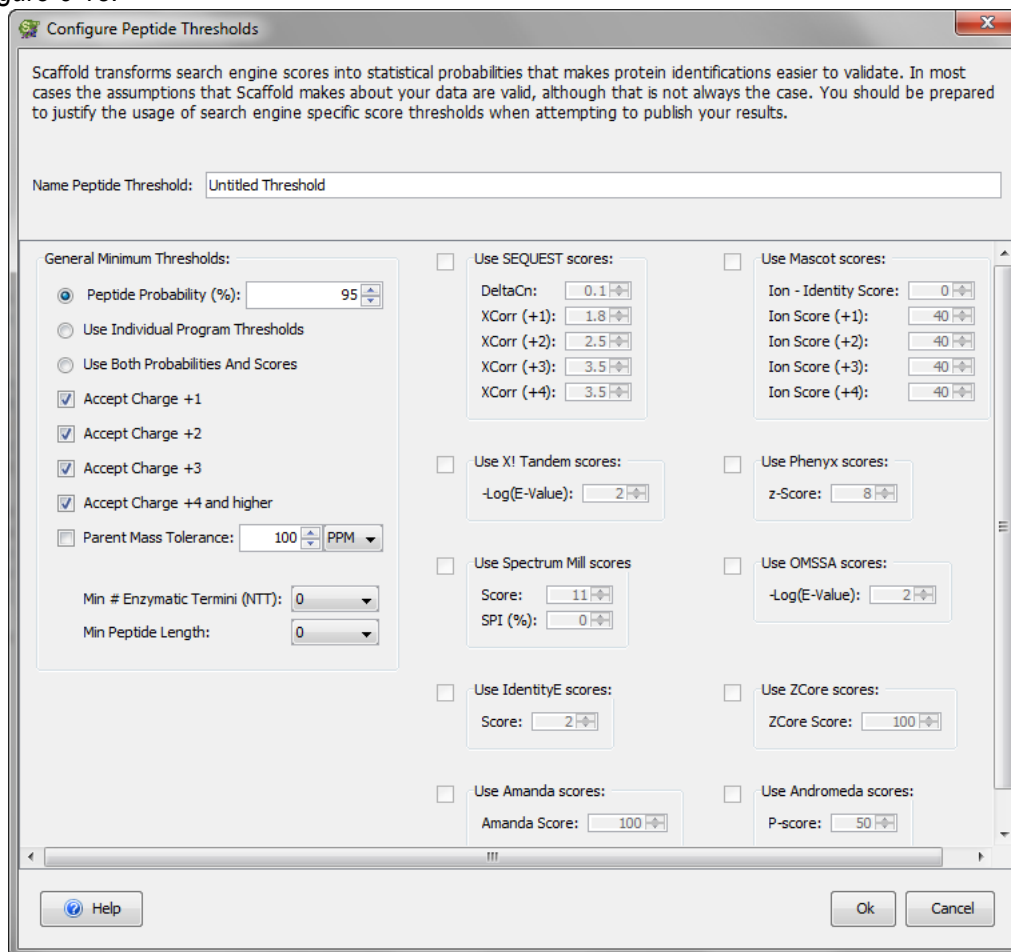


- The filter criteria in effect can be reviewed on the Publish View Page and in the Publications Report.
- Scaffold may work slower for custom filters.
- The ability to define and apply custom filters can also be controlled by a password, which can make sharing correctly displayed datasets between colleagues easier, see Preferences, [Password](#).

Configure Peptide Thresholds Dialog Box

Through the Configure Peptide Thresholds dialog the User can define custom peptide filters to augment the available standard ones.

Figure 6-13:



The following options are available for configuring peptide threshold:

- **Name Peptide Threshold** - Assigns a name to the custom Peptide Probability threshold built in this dialog box.
- **General Minimum Thresholds**
 - *Use Individual Program Thresholds* - Uses only database program information in determining which proteins to display. Choosing this will ignore and disable the protein and peptide probability options.

Note: This is an appropriate option to choose if, in the Statistics View page, the Sequest (XCorr only) distribution histogram displays largely overlapping assigned incorrect and correct matches.

- *Use Both Probability and Scores* - To use both peptide probabilities and search engine scores when filtering data

Note: Unlike the Use Individual Program Thresholds option, this filter does not ignore the Minimum Protein ID Probability.

- *Accept Charges* - Use these check boxes to define which charges Scaffold displays.

- *Parent Mass Tolerance* - The Parent Mass Tolerance is an after-the-probability-calculation filter on the mass accuracy.
- [Min # Enzymatic Termini \(NTT\)](#)
- [Min Peptide Length](#)
- *Program Scores*: These check boxes determine what scores from each search engine filters out the appropriate proteins.



Some Scaffold filtering operations are faster using the standard peptide filters than using the custom peptide filters

Min # Enzymatic Termini (NTT)

When peak lists are searched with a search engine such as Sequest, Mascot, OMSSA, Phenyx, Spectrum Mill or X! Tandem, two of the parameters set are the digestion enzyme and the number of missed cleavages. The search engine only matches spectra to peptides which conform to these parameters.

One approach to increasing the likelihood that the peptides found are correct is to specify that there is no enzyme when running the search engine, and then restricting the search to peptides conforming to the digestion enzyme. Since trypsin is the most common digestion enzyme, the filter in Scaffold is called NTT (Number of Tryptic Termini).

By excluding the peptides with good scores which are non-tryptic, the number of false positives decreases but so does the sensitivity. This filtering on NTT is similar to searching with a loose mass tolerance and then restricting to look at only peptides within a tight mass tolerance. Both approaches are ways to filter the data which are independent of filtering the data on the peptide and protein probabilities calculated by Scaffold.

Min Peptide Length

Filters out peptides with less than the minimum peptide length. This filter can be used to exclude short peptides which are seldom unique to a single protein. These short peptides may cause a very large number of similar proteins to be displayed in the Similarity View.



- *Most search engines (Mascot, Sequest, X! Tandem, etc.) have a minimum peptide length filter option. The Scaffold minimum peptide length filter is only useful if this filtering was not done on the search engine.*
- *Using the minimum filter option on the search engine will reduce the processing and file sizes in Scaffold.*

FDR Filtering

Scaffold allows the User to filter on peptide and/or protein FDR rates when analyzing results of a decoy search. When search results that include decoy matches are loaded in Scaffold, the Peptide and Protein Threshold pull down list includes **%FDR** values in their lists of selectable values. In addition, it is possible to type a custom FDR threshold directly into the box by adding ?FDR? to the end of the string, e.g. ?10.3% FDR?.

FDR filtering in Scaffold works by finding the combination of peptide and protein probability thresholds that maximizes the number of proteins identified without exceeding the FDR thresholds and using the selected minimum number of peptides as a lower bound. An FDR landscape, a matrix with all possible combinations of protein and peptide thresholds, is created and the exact point which maximizes number of proteins while hitting the desired FDR limitations is found. When different threshold combinations would result in the same number of target proteins identified, points at which the protein probability is highest are considered, and of these the point with the highest possible peptide probability is selected.

The actual filtering is then done using the resulting probability threshold settings. The Minimum Peptide Probability and Minimum Protein Probability thresholds selected by the program are shown in the [FDR Dashboard](#), lower left corner of the Scaffold Window. The actual peptide and protein FDR levels are calculated and displayed in the FDR Dashboard as well.

How FDR values are calculated in Scaffold

- Peptide FDR is calculated as the sum of the [Exclusive Spectrum Counts](#) of decoy proteins divided by the sum of the Exclusive Spectrum Counts of target proteins, converted to a percentage.
- Protein FDR is the number of decoy proteins divided by the number of target proteins, expressed as a percentage.

The Display pane

Through the Display pane the User can specify the value (for example, the Number of Assigned Spectra) that is displayed for each protein in each BioSample or MS Sample in the Samples Table. The pane also contains filtering options for limiting the display to only those proteins that meet specific criteria.

Figure 6-14: Scaffold Display pane



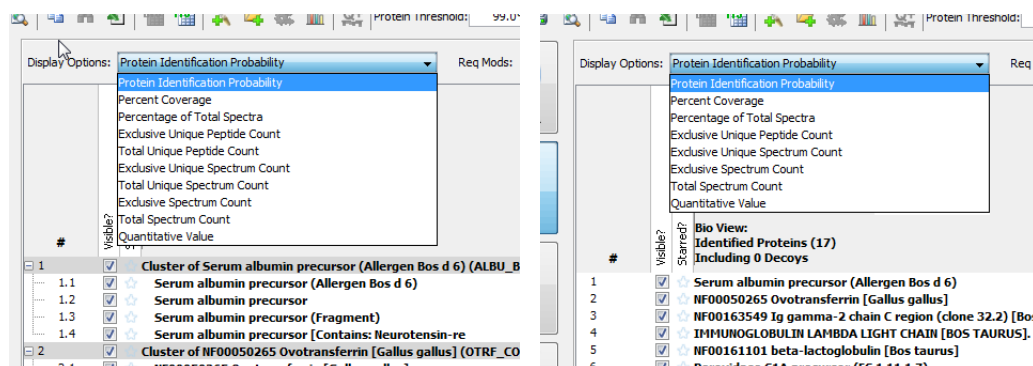
The Display pane contains the following features:

- “Display Options”
- “Req Mods”
- “Search - Text Box”
- “Advanced Search”

Display Options

Scaffold reports statistics other than the identification probability. The Display Options pull down list offers a range of statistics values that once selected are then displayed for each protein under each BioSample or MS Sample in the Samples Table. Depending on whether the clustering algorithm option is selected or not, a slightly different list of options is available.

Figure 6-15: List of Display Options with and without clustering option selected



- **Protein Identification Probability**-- Scaffold’s calculated probability that the protein identification for any of the MS Samples is correct. Results are color-coded to indicate significant differences in protein ID confidence.
- **Percentage Coverage**--The percentage of all the amino acids in the protein sequence that were detected in the sample.

- **Percentage of Total Spectra**-- The number of spectra matched to a protein, summed over all MS Samples, as a percentage of the total number of spectra in the sample.
- **Exclusive Unique Peptide Count** - (corresponding to Number of Unique Peptides in Scaffold3) -- The number of different amino acid sequences, regardless of any modification that are associated with a single protein group or PEG.
- **Total Unique Peptide Count** - (only available with clustering algorithm selected) -- Number of different amino acid sequences that are associated with a specific protein including those shared with other proteins
- **Exclusive Unique Spectrum Count** - (corresponding to Number of Unique Spectra in Scaffold3) -- Number of distinct spectra associated only with a single protein group or PEG. Spectra are considered distinct when they identify different sequences of amino acids or peptides; within the same identifies sequences of amino acids if they identify different charge states or a modified form of the peptide.
- **Total Unique Spectrum Count** - (only available with clustering algorithm selected) -- Number of unique spectra associated with a specific protein including those shared with other proteins
- **Exclusive Spectrum Count** - (corresponding to Number of Assigned Spectra in Scaffold3) -- The number of spectra, associated only with a single protein group or PEG.
- **Total Spectrum Count** - (corresponding to Unweighted Spectrum count in Scaffold3) -
- The total number of spectra associated to a single protein group, or PEG including those shared with other proteins.
- **Quantitative Value** (*Selected quantitative method*) -- Scaffold will display the results of the Quantitative Method selected from the Quantitative Analysis Dialog Box.



When a display option different from Protein Identification Probability is selected, the colored highlights don't change. The colors continue to represent the probability ranges specified by the legend. This is true no matter which statistic is chosen to view, so that a feel for how probable a given identification is, is always available.

Req Mods

The Required Modifications filter lists all the post-translational modifications (PTMs) selected during the search phase of data processing. Choose a modification on the drop-down list to filter the display to only those proteins, peptides, and spectra that contain the selected modification.

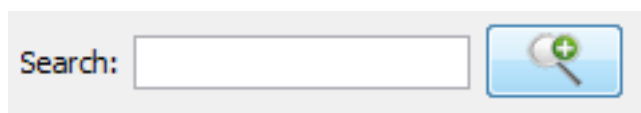
- **No Filter** - No filtering is applied. All proteins, peptides, and spectra that meet all other display and filtering options are displayed.

- **Unmodified Only** - Display only those proteins, peptides, and spectra that do not have any associated PTMs.
- **Variable Modifications** - Display only those proteins, peptides, and spectra that were identified as having the selected variable modification

Search - Text Box

The Scaffold search box allows the user to type in search terms to quickly identify specific proteins by protein names or accession numbers, but it can also filter on peptide sequences and/or spectra information.

Figure 6-16: Search text box



The Search field accepts regular expressions and filters the results based on accession number or protein name. Only those proteins that meet all the search criteria are displayed. Your search is limited to the exact order of the characters in the string, but the string is not case-sensitive and it can appear anywhere in the search results. For example, a search string of **ATP** returns both **ATP** synthase and calcium-transporting **ATP**ase. In another example, a search string of **sodium|transport** returns all values that have sodium and/or transport in the protein name - **Sodium**/potassium **transporting**... and Calcium-**transporting** ATPase sar..., and so on.

Click on the magnifier glass button to the right side of the search text box for more advanced search features. See [“Advanced Search”](#)

Advanced Search

When clicking on the magnifier glass button on the right hand side of the search text box in the Samples View the Configure Advanced Protein Filter dialog opens.

Figure 6-17: Samples View - Configure Advanced Protein Filter

The dialog contains a number of tools the User can use to search for specific proteins, peptides, spectra (useful in peptidomic studies whenever questions arise about a peptide assignment) or peptide motifs (useful to investigate potential modification sites).

Searches can be performed over the full list of identified proteins or within different groups of proteins like categories or starred proteins.

The presence/absence search options allow for searches based on the intersection of categories. It displays only proteins found in a category, or proteins found in one category and not in another category.



The presence/absence search works in a similar manner as Scaffold's Protein Venn Diagram. This feature requires the use of regular expressions.

Searches can be performed also on Taxonomy and GO terms after they are added to the Samples list.

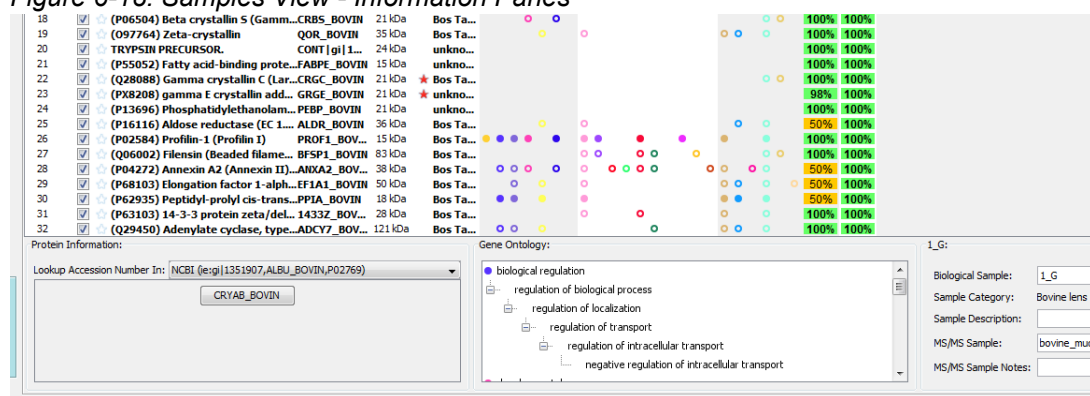
Information Panes

The bottom section of the Samples View contains three information panes:

- “Protein Information pane”
- “Gene Ontology pane”
- “Sample Information Pane”

Each pane provides further diversified information related to each row in the Samples Table

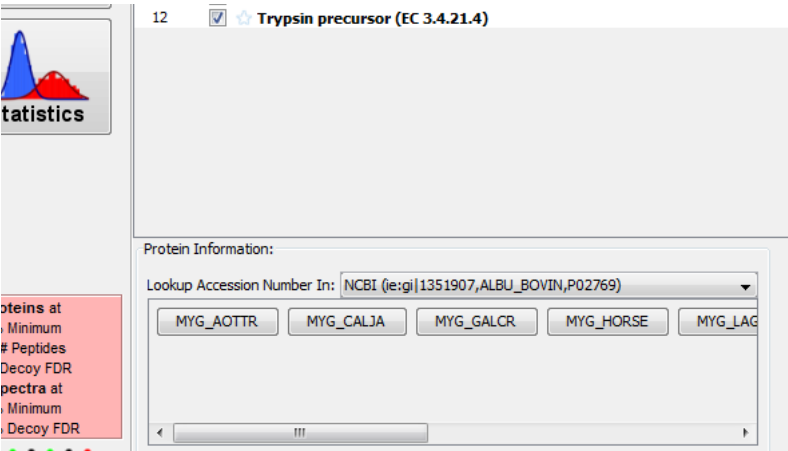
Figure 6-18: Samples View - Information Panes



Protein Information pane

The Protein Information pane is displayed in the lower left section of the Samples View. It includes a look up accession number pull down list of Online protein databases such as SwissProt or NCBI. For each selected row in the Samples Table, the pane shows the set of proteins included in the corresponding protein group as click-able buttons. Clicking one of the buttons opens an Internet browser to the address selected from the pull down list and searches for the selected protein accession number. If the accession number is found additional information for the selected proteins is then easily available to the user.

Figure 6-19: Protein Information pane

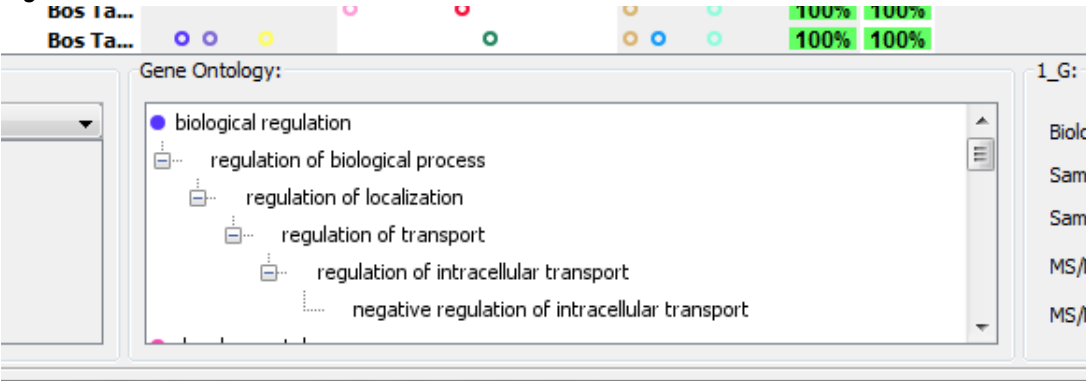


Gene Ontology pane

The Gene Ontology pane when displayed is located in the lower center section of the Samples View. The pane is displayed only whenever the GO terms have been searched. GO terms are added to the Samples Table either when searched during the loading phase or after the data is loaded by going to the **Experiment> Add Go Annotations**. For more information see [Edit GO Term Options](#).

The terms are displayed structured as a term ancestry, with the high-level GO annotations showing as colored dots (which match the colors shown in the Samples Table) and its subsequent children.

Figure 6-20:





- Double-clicking a GO term in this pane opens a page in a browser with detailed information about the term.
- GO terms may be hidden by un-checking the menu in **View > Show GO Annotations**. See [Edit GO Term Options](#)

Sample Information Pane

The Samples Information Pane displays the Biological and MS Sample names and descriptions for the selected MS Samples. Biological Sample name and notes, and the MS Sample name and notes can be edited here. To populate this pane the User needs to click on a Bio/MS Sample column.



To change a category name for a BioSample, go to **Experiment > Edit BioSample**. See [Edit BioSample](#).

Chapter 6

The Samples View

Chapter 7

Quantify View

Scaffold's Quantify View provides graphical tools to help the User visualize experiments and draw conclusions about the quantitative relationships demonstrated in the data. From the Quantify View, the User can compare spectral counts between samples and categories, analyze the biological functions of the proteins identified in the experiment, and assess the reliability of the statistical analysis of the data.

This chapter details the Scaffold's Quantify View including the information that can be gleaned from this view as well as the features of the view.

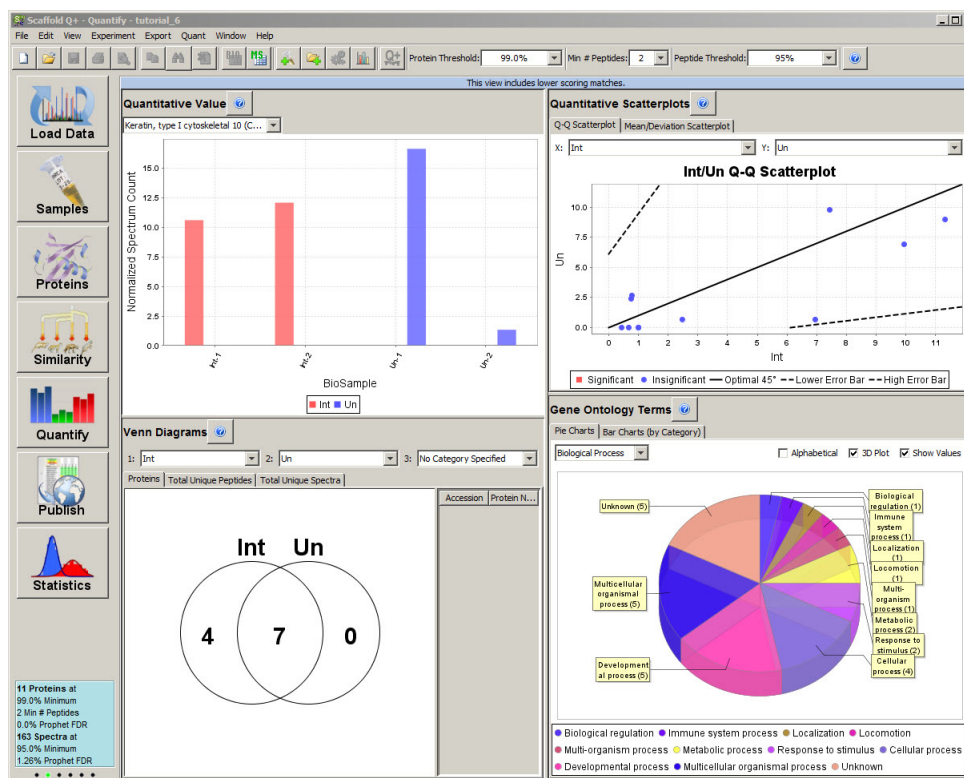
This chapter covers the following topics:

- [“The Quantify View” on page 144.](#)

The Quantify View

The Quantify View provides overviews and tools to help the User make direct comparisons among BioSamples or MS Samples. It can be reached through the Quantify button located in the [Navigation pane](#) or through the menu **Window > Quantify**.

Figure 7-1: Scaffold Quantify View



The Quantify view includes the following panes:

- The [Quantitative Value](#) pane, in the upper left of the Quantify View, provides information about relative quantities of a specific protein and allows comparisons between biosamples and categories.
- The [Quantitative Scatterplots](#) pane, in the upper right of the Quantify View, shows the degree of error associated with the spectral count measurements.
- The [Venn Diagrams](#) pane, in the lower left of the Quantify View, shows the relationship between proteins, total unique peptides and total unique spectra in various categories, and allows the User to easily identify proteins, peptides or spectra of interest.
- The [Gene Ontology Terms](#) pane, in the lower right of the Quantify View, helps identify which proteins may be biologically significant

The Quantitative Value pane

The Quantitative Value pane contains the Normalized Spectrum Count bar chart and located above the chart, a pull down list of the proteins appearing in the Samples table.

The Normalized Spectrum Count bar chart provides a view of the relative abundance of a specific protein (selected through the pull down list) across different BioSamples and categories.

- The Y-axis displays the normalized count of the spectra matching any of the peptides in the selected protein. This count depends upon the protein, peptide, required mods, and search filters and thresholds set on the [The Samples View](#).
- X-axis displays bars for each BioSample in the Scaffold experiment. The bars are color coded according to the defined categories.

If the loaded dataset contains replicates, from this pane the User can assess the consistency of spectral counts across replicates within each category while comparing expression levels of the protein between categories. This allows visual inspection of the data and provides insight into the meaning of statistical comparisons such as the T-test or ANOVA.

The Quantitative Scatterplots pane

The Quantitative Scatterplots pane includes two tabs:

- [Q-Q Scatterplot tab](#)
- [Mean/Deviation Scatterplot tab](#)

Q-Q Scatterplot tab

This tab contains a scatter plot and two pull down menus used to assign categories to each of the axis in the plot. The Q-Q Scatterplot helps evaluate the degree of error associated with the spectral count measurements. The graph plots each protein appearing in the [Protein List](#) as a point on a two dimensional scatter plot. If the categories have multiple replicates, then the average value in the category is plotted:

- **X-axis:** normalized spectral count for the protein for all samples in a category.
- **Y-axis:** normalized spectral count in a second category.

If a Quantitative Analysis Test has been applied through the [Quantitative Analysis...](#) command, the proteins labeled significantly different by the currently applied test between categories are plotted as red points. All other proteins plot as blue points. Any two categories of proteins can be chosen for display for the X and Y axes from the drop down lists.

The plot includes the following functionalities:

- Hovering over a point displays a tool-tip identifying the protein represented by that point and giving its precise coordinates.
- Double-clicking on a point takes the User to the Proteins view for that protein.

- Scaffold draws a line with a slope of 1 on the graph. Proteins with similar abundances in both categories should plot as points near this line. There are also two dashed lines drawn on the figure. Proteins that plot outside these lines are more than two standard deviations away from being the same in both categories. These proteins are differentially expressed. The error lines are estimated from the [Mean/Deviation Scatterplot tab](#).

Tip: The two standard deviation lines diverge from the 45 degree line because the standard deviation depends upon the number of spectra.

Mean/Deviation Scatterplot tab

The Mean/Deviation Scatterplot provides a method of estimating the coefficient of variation (or variance) (CV) of the estimates of protein abundance. The tab includes a graph that plots the mean and standard deviation for each protein appearing in the protein list for the whole data set.

- X-axis: average or mean value of the estimated protein abundance across all samples
- Y-axis: standard deviation of the estimated protein abundance computed across all samples.

A regression line is calculated to provide a model that defines the two standard deviation lines shown in the plot included in the [Q-Q Scatterplot tab](#). This theoretical estimation is represented as a dashed line in the plot and shows in general that the larger the estimated protein abundance, the larger is the absolute uncertainty in the estimate.

Another way of using this graph is to evaluate if the percent uncertainty, the CV, is roughly constant see reference [Pavelka \(2008\)](#). This method of estimating the CV uses all the available data. In most instances using all the data for an estimate gives the best estimate. However like any time a line is fit to data, it is possible for outliers to cause inaccuracies. Outliers in this data that will introduce the most inaccuracies are proteins with a high estimated abundance in the samples of one category and low abundance in another sample category. The implicit assumption is that if some abundant proteins are greatly suppressed in one category of samples, they will be balanced by roughly the same number of abundant proteins with elevated levels. If a visual inspection of this graph suggests that outliers have distorted the estimate of the CV, then care should be taken when interpreting the Q-Q Scatterplots.

The Venn Diagrams pane

The Venn Diagram pane includes three tabs each containing Venn diagrams with different types of numbers and a table placed on the right side of the diagram. The table becomes visible whenever the User selects a section of the diagram. When a section of the diagram is selected it appears highlighted in yellow. Over the tabs there are three drop down lists showing the categories available in the experiment.

Through the Venn Diagram pane, the User can take a look at the relationship among proteins, total unique peptides, or total unique spectra identified in the various categories.

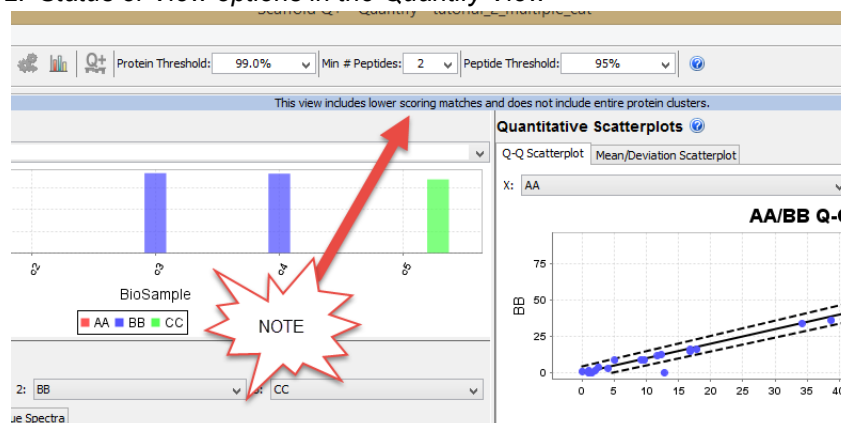
Each of the tabs display a Venn diagram showing the overlap of up to three categories and reflecting the current filters and thresholds applied in the Samples View. The User can determine which category is visible in the diagram through the drop down lists.

- **Proteins tab** - The diagrams show the number of proteins identified in each category and in the overlap between up to three categories. When a section of the diagram is selected, the table shows the list of proteins included in the highlighted section.
- **Total Unique Peptides tab** - The diagrams show the sum of [Total Unique Peptide Count](#) for each protein in a category for up to three categories. When a section of the diagram is selected, the table shows the list of peptides included in the highlighted section.
- **Total Unique Spectra** - The diagrams show the sum of the [Total Unique Spectrum Count](#) for each protein in a category for up to three categories. When a section of the diagram is selected, the table shows the list of identified peptides, with their charge and modifications, included in the highlighted section.

The numbers shown in the Venn diagrams include all proteins and peptides displayed in the Samples view.

- If the View option Show Entire Protein Clusters is selected, the counts will include the lower-scoring proteins that are part of displayed clusters, see [Clusters in the Samples Table](#).
- if the View option [Show Lower Scoring Matches](#) is on, the numbers include peptides not meeting the current thresholds but currently displayed because they do meet the thresholds in other biosamples for the same protein. The status of these options is shown at the top of the window. To count only proteins and peptides that meet thresholds, the User can go to the View menu and turn off these options.

Figure 7-2: Status of View options in the Quantify View



•



- *The Venn Diagram is interactive. When selecting a region of the diagram, the protein accession numbers, peptide sequences, or spectral peptide sequence and nominal charge (spectra display) display in the table next to the Venn diagram.*
- *When double clicking on a region of the Venn diagram, Scaffold switches to [The Samples View](#) and applies an Advanced Filter so that only those proteins in the selected region are displayed. The **Search** input box becomes highlighted in yellow and displays **Advanced...***
- *To remove the Venn Diagram's applied advanced filter the User needs to clear the contents of the highlighted yellow input box or simply double-click outside the Venn Diagram in the diagram pane.*

Gene Ontology Terms pane

The Gene Ontology Terms pane gets populated only when GO terms have been searched and found, see [Apply GO Annotations/ Apply NCBI/ Configure GO annotation Sources](#) and [Edit GO Term Options](#). When the terms have been added each protein displayed in the Samples Table may show one or many Gene Ontology Terms describing it. These terms are very useful to attach biological significance to the results.

The detailed GO terms describing each protein are summarized in the pane in broader categories called ontologies. Each one of these ontologies has its own pie and bar charts. The User can select which ontology to display using a drop down above the chart. The three ontologies categories available are:

- Biological process
- Cellular component
- Molecular function
-

Pie Charts

Each slice of the pie chart corresponds to one column of the GO term annotations in the Samples view. The GO term represented by a slice is shown in a box linked to the slice. If Show Values is checked the number of proteins annotated with that GO term is also shown. Since a single GO term may be associated with more than one protein, these numbers may sum to a value greater than the number of proteins.

Double-clicking on a section of the pie chart filters the proteins in the experiment to show only the proteins with annotated with that GO term and brings up the Samples view. A filtered set can be further filtered by returning to the pie chart and double-clicking again. More sophisticated GO filtering can be done through the Advanced Filters dialog.

Bar Charts

The bar charts are organized by category. Each bar displays the number of proteins annotated with a certain GO term in a certain category. This allows you to compare if proteins associated with a certain biological structure or function are differentially expressed in one category or another.

Chapter 7
Quantify View

Chapter 8

Protein Grouping and Clustering

This chapter describes the way Scaffold groups and thins out the list of proteins shown in the Samples Table, so the User can focus on the most likely protein identifications present in the experiment. The grouping and paring is achieved using different types of algorithms depending on whether the option **Protein Cluster Analysis** is selected or not.

The chapter details the different grouping and clustering algorithms used in Scaffold as follows:

- [“Shared Peptide Grouping and Protein Cluster Analysis” on page 152](#), which provides a description of the Shared Peptide Grouping algorithm used for grouping, of the paring and clustering of proteins appearing in the Samples Table for version 4 and higher.
- [“Legacy Protein Grouping” on page 159](#), which provides a description of the grouping algorithm used in versions 3 and older and still applied in Scaffold version 4 and higher when the clustering option is not selected.

Shared Peptide Grouping and Protein Cluster Analysis

Scaffold version 4 and higher includes the option of applying a method of grouping proteins called Shared Peptide Grouping. Scaffold versions 3 and lower instead used a different grouping algorithm referred to as the Legacy Protein Grouping.

Shared Peptide Grouping is designed to lessen the probability of discarding a valid protein identification when the protein happens to share many peptides with another identified protein. Scaffold version 4 and higher also includes the option to assemble proteins into clusters based on shared peptide evidence using Protein Cluster Analysis.

These two options are selected during file loading by checking **Use protein cluster analysis** in the **Load and Analyze** Wizard page within the **Protein Grouping** pane. Choosing this option enables the application of both Shared Peptide Grouping and Protein Cluster Analysis.

Figure 8-1: Load and Analyze Window

Load and Analyze Data

Searched Database:
uniprot_sprot_mouse_20121129 FASTA Database (2)
☐ Use non-default forward/decoy ratio: No Decoys
Add New Database

X! Tandem:
☐ Analyze with X! Tandem

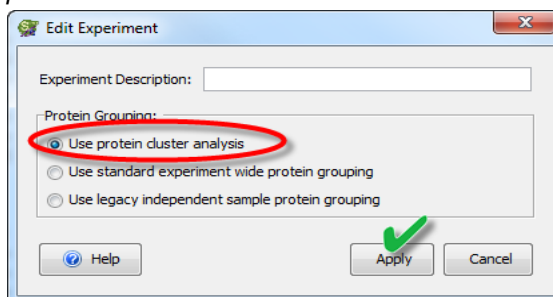
Scoring System:
☒ Use LFDR scoring (all instruments)
☐ Use legacy PeptideProphet scoring (high mass accuracy)
☐ Use legacy PeptideProphet scoring (standard)

Protein Grouping:
☒ Use protein cluster analysis
☐ Use standard experiment wide protein grouping
☐ Use legacy independent sample protein grouping

Protein Annotations:
☒ Don't annotate (No download required)
☐ Fetch GO annotations remotely (UniProt, IPI, NCBI; ~20 mins every time)
(Configure GO Source)

If during file loading Protein Cluster Analysis is not selected, it can be reapplied to the already loaded data by going to the menu **Experiment> Edit Experiment**. The **Edit Experiment Window** opens and, in the full version of Scaffold, the check box **Use Protein Cluster Analysis** is available. Selecting the check box and clicking **Apply** rearranges the protein groups and creates clusters using Shared Peptide Grouping and Protein Cluster Analysis.

Figure 8-2: Edit Experiment Window



For explanation purposes the grouping and clustering processes can be broken down into the following three phases:

- Protein Grouping
- Protein Paring
- Protein Clustering

Protein Grouping

The way Shared Peptide Grouping assigns peptides to proteins is quite different from how it is done in the Legacy Protein Grouping algorithm. Rather than assigning each peptide to a single protein, Shared Peptide Grouping includes a peptide in all of its matching proteins. It then precedes to form “Protein Groups” and assign weights to each shared peptide, see “Weighting Function”

Protein Groups

Scaffold considers proteins that share peptide evidence. In cases where two or more proteins share all of their peptides, there is no basis for discrimination amongst them and the proteins are grouped and treated as a unit called *Protein Group*. These proteins appear in the Samples Table as a single line with the accession number of one of them followed by a plus and the number of other proteins in the group. The “preferred” or named protein is arbitrarily selected and may be changed by the user.

Figure 8-3: Samples Table, Protein grouping

Cytochrome b-c1 complex subunit 8 OS=Mus musculus GN=U...	QER8_MOUSE	10 kDa	LK3 transgenic mice	● ●
40S ribosomal protein S29 OS=Bos taurus GN=RPS29 PE=3 SV...	R529_BOVIN (+6)	7 kDa	Bos Taurus	● ●
Hornerin OS=Homo sapiens GN=HRNR PE=1 SV=2	HRNR_HUMAN	282 kDa	Homo sapiens	● ●

Weighting Function

For the purpose of calculating the protein probabilities, shared peptides are apportioned among proteins according to a weighting function.

The weights are assigned by using the following formula:

$$W(p, A) = \frac{PE_{excl}(A)}{\sum_{All(B \supseteq p)} PE_{excl}(B)}$$

Where $W(p, A)$ is the weight assigned to shared peptide p contained in protein A and in other proteins. $PE_{excl}(A)$, the exclusive peptide evidence, is defined as the sum of the probabilities of each exclusive unique valid peptide X belonging to protein A .

$$PE_{excl}(A) = \sum_{X \subset A} P_X$$

This value is then normalized by the sum of the exclusive peptide evidence for each of the proteins that contain peptide p .



*A peptide can be set “valid” either manually, by un-checking peptides in the Proteins View Peptide Table or globally by using the Experiment menu option **Reset Peptide Validation**. The Scaffold default cut off is 0%.*

As an example [Figure 8-4](#) shows peptides shared by multiple proteins with their related weights listed in the Similarity view.

Figure 8-4: Similarity View- Peptide weights

Myoglobin									
Index	Peptide	Prob	Exclusive To	Valid	Cluster of Myoglobin (MYG_HORSE)				Cluste... No Group
					MYG_HORSE (+1)	MYG_CA5FI	MYG_GALCR (+2)	MYG_OCHPR	MYG_ORYAF
1	ADIAGHGQEVLR	100%		<input checked="" type="checkbox"/>	0.76	0.02		0.11	
2	ALELFR	33%		<input checked="" type="checkbox"/>	0.63	0.01	0.09	0.09	0.09
3	ETLEKFDKFNLIKSEDEMKGS...	100%	Myoglobin	<input checked="" type="checkbox"/>				1.00	
4	GDFGADAQGAMTK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00				
5	GLSDGEWQVVLNWWGK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00				
6	HGTVVLTALGGILK	100%	Myoglobin	<input checked="" type="checkbox"/>	1.00				

These weights are also displayed in the Proteins View Peptide Table in the column titled “Apportionments,” which replaces the traditional “Assigned” column when the cluster grouping model is used.

Protein Paring

Next, the protein list is pared down according to the principle of parsimony. As in the case of

the Legacy Protein Grouping, the Shared Peptide Grouping algorithm thins down the list of proteins by eliminating any for which there is no independent evidence. However, independent evidence is defined differently in the two grouping algorithms. In the Shared Peptide Grouping a protein is considered having independent evidence when it contains at least one exclusive unique peptide.

Proteins for which there is no exclusive evidence are then eliminated from the protein identification list. This process can best be seen in Scaffold's Similarity View. Here all proteins sharing peptide evidence are assembled into a table. Proteins with exclusive peptides are placed to the left and included in the experiment. Proteins for which all of the associated peptides are subsumed by these identified proteins are eliminated from further consideration, as there is no independent evidence of their presence in the experiment. These proteins appear in the "No Group" columns in the Similarity View but are otherwise invisible in Scaffold.

Figure 8-5: Similarity View - No group columns

Histone H2A type 1-B/E OS=Homo sapiens GN=HIST1H2AB PE=1 SV=2

Index	Peptide	Prob	Exclusive To	Valid	Cluste... H2A1B_HUMAN (+14)	Cluste... H2A1D_HUMAN (+18)	Cluste... H2AV_BOVIN (+21)	Cluste... H2AZ_CANAL (+1)	No Group H2A1A_HUMAN (+25)	No Group H2A1_CANAL (+30)	No Group H2A1_ASHGO (+40)
1	ATIAGGGVIPHIK	52%	Histone H2...	<input checked="" type="checkbox"/>			1.00				
2	HLQLAIR	80%		<input checked="" type="checkbox"/>	0.39	0.39	0.21	0.01			
3	HLQLAIRNDEELNK	100%		<input checked="" type="checkbox"/>	0.50	0.50					
4	NDEELNKLKG	100%	Histone H2...	<input checked="" type="checkbox"/>		1.00					
5	NDEELNKLGR	100%	Histone H2...	<input checked="" type="checkbox"/>	1.00						
6	VTIAQGGVLPNIQAVLLPK	100%		<input checked="" type="checkbox"/>	0.50	0.50					

Protein Clustering

Assembling proteins into clusters is based on shared peptide evidence. While akin to Mascot's hierarchical family clustering, Scaffold's Protein Cluster Analysis is more stringent in its requirement for two proteins to appear in the same cluster. This added stringency often succeeds in separating proteins into sets of biologically meaningful isoforms.

In essence, a cluster is a set of proteins with overlapping peptide evidence, and may be treated as a proxy for a single identification. This view allows interpretation of identification probability, spectral counts, and normalized quantitative values calculated on the level of clusters.

Cluster formation begins with the creation of protein groups as described above, see "[Protein Groups](#)". Next, these protein groups are grouped into clusters of similar proteins. Two proteins (or protein groups) are considered similar if their joint weighted peptide evidence is at least half of the weighted peptide evidence of either protein. A protein is iteratively added to a cluster if it is similar to at least one other protein in the cluster. This information can be translated into the following rules of thumb for cluster formation:

1. For two proteins to be clustered, the sum of the probabilities of their shared peptides must be at least 95%.

2. The proteins must share at least 50% of their evidence. This is determined by summing the probabilities of the shared peptides and comparing this value with the summed probabilities of all of the peptides for each individual protein. If the sum of the probabilities of the shared peptides is greater than or equal to half of the sum of the peptide probabilities for either of the individual proteins, a cluster is formed.
3. A protein may be included in an existing cluster if it meets the above criteria with a member protein of the cluster.

For a detailed example of how a cluster is formed see an extended version of this document published on our website: [scaffold_protein_grouping_clustering.pdf](#).

Clusters in the Samples Table

Thresholds and filters do not affect the formation of clusters, but they do determine which clusters and proteins or protein groups are displayed in the Samples Table. Scaffold builds the Samples Table applying thresholds and filters to the formed clusters, proteins and proteins groups in the following order:

1. Select all clusters that pass thresholds.
2. Include all proteins and protein groups belonging to the selected clusters.
3. Prune proteins or protein groups based on selected filters
4. Remove clusters that do not include proteins
5. Prune proteins and proteins groups based on thresholds.

This order of applying thresholds and filters keeps clusters in the Samples Table that might not include proteins or protein groups that pass thresholds and filters.



Filters apply only to proteins or protein groups.

Clusters are shown in the Samples View as a line with protein name “Cluster of ...” and the name of one of the constituent proteins. This protein is designated as the “primary protein” of the cluster, but the primary protein may be changed by clicking on the accession number field of a cluster and selecting a different accession number from the drop-down list when it appears. A cluster may be expanded in the Samples View by clicking on the “+” at the left of the cluster’s row. When a cluster expands, it displays all of its constituent proteins or protein groups, including the primary protein. The right click menu provides bulk operations to expand or collapse all clusters simultaneously.

The menu option **View > Show Entire Protein Cluster** will show in a gray font proteins or protein groups in clusters that do not pass thresholds

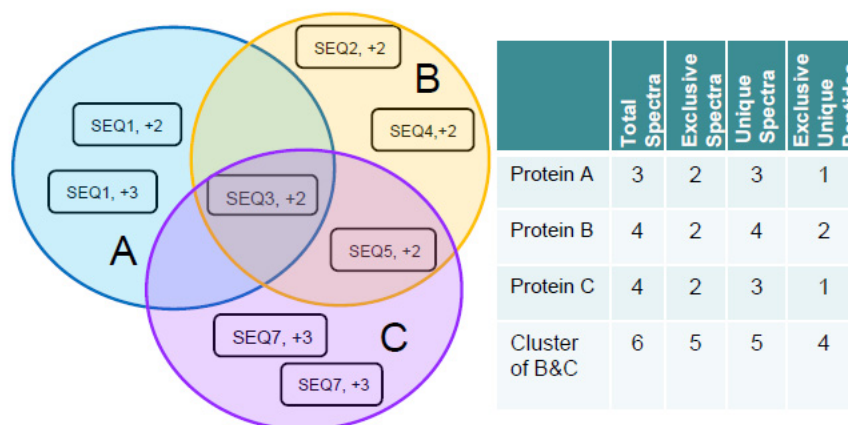
Clusters Display Values

Display values are calculated for a cluster as a whole based on the set of peptides that make

up the cluster. Note that these values are different from the values of any individual protein, including the primary protein of the cluster. Selecting a cluster and going to one of the other views displays all of the information for the entire cluster.

Diagram in [Figure 8-6](#) illustrates Scaffold's method of spectra and peptide counting in clustered proteins. The circles A, B and C represent three proteins of which B and C form a cluster. The little squares in the circles represent the spectra included in the proteins. Their charge is also indicated. The table on the side shows how the different quantitative values are counted for each protein and for the cluster. Note that the total spectra of the cluster does not correspond to the sum of the total spectra of the proteins included in the cluster because some of the peptides are shared.

Figure 8-6: Spectral counting in clustered proteins.



Legacy Protein Grouping

Scaffold 4 groups proteins with the Legacy Protein Grouping algorithm used in its versions 3 and older when, during the loading phase, the clustering option is not selected or when data is already loaded in Scaffold and the option **Use Protein Cluster Analysis** is not selected in **Edit Experiment Window**.

Generally, the Legacy Protein Grouping algorithm groups proteins using a table very similar to the one shown in the Similarity View when the Protein Cluster Analysis option is not selected, see [Figure 8-7](#).

Figure 8-7: Scaffold Legacy Similarity View

The screenshot displays the 'Similarity' tab in Scaffold. The table is titled 'Serum albumin precursor (Allergen Bos d 6)'. It lists 17 peptides (Index 1-17) and compares them against 13 protein groups: ALBU_BOVIN, ALBU_CONTR, *ALBU_PIG, ALBU_RABIT, *ALBU_RAT, ALBU_SHEEP, ALBU_MACMU, ALBU_CANFA, ALBU_FELCA, ALBU_HUMAN, ALBU_HORSE, ALBU_MOUSE, and FETA_HORSE. The 'Valid' column shows checkmarks for all peptides. The 'Exclusive To' column lists 'Serum albu...' for peptides 1-6 and 15-17. The probability values are shown in the cells, with some highlighted in green (100%), yellow (62%), or red (29%, 26%, 36%, 40%).

Index	Peptide	Exclusive To	Valid	ALBU_BOVIN	ALBU_CONTR	*ALBU_PIG	ALBU_RABIT	*ALBU_RAT	ALBU_SHEEP	ALBU_MACMU	ALBU_CANFA	ALBU_FELCA	ALBU_HUMAN	ALBU_HORSE	ALBU_MOUSE	FETA_HORSE
1	AADKDNCFATEGPNLVARSK...	Serum albu...	✓					95%								
2	AATITK		✓													(36%)
3	AEFVEVTK	Serum albu...	✓	100%	100%											
4	AIPEPLPLTADFAEDKDVCK	Serum albu...	✓	100%	100%											
5	CCAADDOKEACFAVEGPK	Serum albu...	✓	100%	100%											
6	CCTESLVNR		✓	100%	100%	(100%)			(100%)	(100%)		(100%)	(100%)			
7	CDNQOTISSK	Serum albu...	✓	62%	62%											
8	DAFLGSFLYEYSR	Serum albu...	✓	100%	100%											
9	DAIPENLPLTADFAEDKDVCK	Serum albu...	✓	100%	100%											
10	DOPHACYSTVFVK	Serum albu...	✓	100%	100%											
11	DLGEEHFK	Serum albu...	✓	100%	100%					(100%)						
12	ECCDKPLLEK		✓	100%	100%	(100%)	(100%)			(100%)						
13	ECHGDLLECADDR		✓	100%	100%	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
14	ECHGDLLECADDRADLAK		✓	100%	100%	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
15	ENFVAFVDK	Serum albu...	✓	29%	29%			(29%)								
16	ETYGDM	Serum albu...	✓	26%	26%			(26%)								
17	ETVGMANCCFK	Serum albu...	✓	40%	40%			(40%)								

Assigning peptides to proteins

Initially the table of peptides and proteins has a column for every protein to which a peptide could potentially be assigned, and a row for every valid peptide that can be found in the listed proteins. When a peptide is found in a protein the peptide probability is shown in the appropriate cell. The sum of the probabilities is then calculated for each protein, see [Figure 8-8](#).

Figure 8-8: Initial Similarity Table

Index	Peptide	Exclusive To	Valid	E	F	G	H	I	J	K	L	M	N	O
1	AKWYPEVR		FALSE	9%	9%	9%	9%	9%			9%	9%	9%	
2	CVVVGDAVGK		FALSE	28%	28%	28%			28%		28%			
3	DDKDTIEK		TRUE	73%	73%	73%	73%	73%				73%	73%	73%
4	GSPQAIK	Chain A, Small G-Protein	TRUE	75%										
5	IISAMQTIKCVVVGDAVGK		TRUE											
6	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%			95%	95%	95%	95%	95%	95%	95%	
7	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%								
8	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%			95%	95%	95%	95%	95%	95%	95%	
9	LVPITYPQGLAMAK		TRUE											
10	TVFDEAIR		TRUE	95%	95%	95%	95%	95%	95%	95%	95%			95%
11	VDSKPVNLGLWDAGQEDYDR		TRUE											92%
12				433%	263%	263%	358%	358%	285%	285%	285%	263%	263%	260%

Each peptide is then assigned to the protein that has the highest total probability among all those where the peptide is found, see Figure 8-9. If two or more proteins have equal total probabilities and that is the highest for that peptide, it is assigned to all of them.

Figure 8-9: Assigned peptides are shown in green, unassigned in gray

Index	Peptide	Exclusive To	Valid	E	F	G	H	I	J	K	L	M	N	O
1	AKWYPEVR		FALSE	9%	9%	9%			9%	9%	9%	9%	9%	
2	CVVVGDAVGK		FALSE	28%	28%	28%	28%		28%	28%	28%			
3	DDKDTIEK		TRUE	73%	73%	73%				73%	73%	73%	73%	73%
4	GSPQAIK	Chain A, Small G-Protein	TRUE	75%										
5	IISAMQTIKCVVVGDAVGK		TRUE											
6	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%	
7	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%				95%	95%			
8	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%	95%	95%	95%	95%	95%			95%	95%	
9	LVPITYPQGLAMAK		TRUE											
10	TVFDEAIR		TRUE	95%	95%	95%	95%	95%	95%	95%	95%			95%
11	VDSKPVNLGLWDAGQEDYDR		TRUE											92%
12				433%	358%	358%	285%	285%	285%	263%	263%	263%	263%	260%

Defining protein Groups

Now the grouping begins. Proteins with no peptides assigned are eliminated from consideration, the evidence for those proteins has already been accounted for in proteins which are more likely to be present in the analyzed sample. Proteins with the same peptides assigned to them are combined into a group, see Figure 8-10.

Figure 8-10: Protein groups formation

Scaffold Table Export.xls															
	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1				Group 1	Group 2	Group 3									
2	Peptide	Exclude To	Valid	#115781969	#113279011	#130585149	#1169158861								
3	AKWYPEVR		FALSE	9%	9%	9%									
4	CVVVGDAVGK		FALSE	28%	28%	28%									
5	DDKDTIEK		TRUE	73%	73%	73%	73%								
6	GSPQAIK	Chain A, Small G-Protein	TRUE	75%											
7	IISAMQTIKCVVVGDAVGK		TRUE												
8	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%											
9	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%									
10	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%											
11	LVPIPTYPQGLAMAK		TRUE												
12	TVFDEAIR		TRUE	95%	95%	95%	95%								
13	VDSKPVNLGLWDTAGQEDYDR		TRUE				92%								
14				433%	263%	263%	260%								

There is one further complication, however. If the only evidence for a group is a single protein with probability less than 95%, Scaffold disregards this group. This is based on a heuristic rule built into the algorithm which cuts down on the number of false protein matches displayed. In this case it would eliminate Group 3, see [Figure 8-11](#).

Figure 8-11: Formed protein groups

Scaffold Table Export.xls															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1					Group 1	Group 2									
2	Index	Peptide	Exclude To	Valid	#115781969	#113279011	#130585149								
3	1	AKWYPEVR		FALSE	9%	9%	9%								
4	2	CVVVGDAVGK		FALSE	28%	28%	28%								
5	3	DDKDTIEK		TRUE	73%	73%	73%								
6	4	GSPQAIK	Chain A, Small G-Protein	TRUE	75%										
7	5	IISAMQTIKCVVVGDAVGK		TRUE											
8	6	KLTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%										
9	7	LIPITYPQGLAMAK	Ras-related C3 botulinum tox	TRUE		95%	95%								
10	8	LTPITYPQGLAMAK	Chain A, Small G-Protein	TRUE	95%										
11	9	LVPIPTYPQGLAMAK		TRUE											
12	10	TVFDEAIR		TRUE	95%	95%	95%								
13	11	VDSKPVNLGLWDTAGQEDYDR		TRUE											
14					433%	263%	263%								

Generally, this approach works well to eliminate false assignments, however in certain instances, it can result in a protein that may actually be found in the sample being eliminated from consideration, and thus not seen in Scaffold's other views. Unfortunately, changing the filter settings has no effect upon this type of grouping algorithm. A different approach can now be tried by using the clustering option available in Scaffold 4. The new grouping algorithm does not forcefully assign peptides uniquely to a protein but considers shared peptides among different proteins.

Chapter 9

Quantitative Methods and tests

Scaffold supports label free quantitative methods. Some of them are based on spectrum counting and others are based on MS1 intensity measurements. For the purpose of establishing differential expressions among the categories present in a Scaffold experiment it is important to normalize the values and accommodate for systematic differences and experimental errors. Scaffold provides option for normalizing values and taking care of missing values.

This chapter describes the various label free quantitative methods available in Scaffold and how Scaffold normalizes them. It also describes the different Quantitative tests available in the program.

- [“Label Free Quantitative Methods” on page 164](#), which describes the quantitative methods available in Scaffold and how they are computed
- [“Normalization among samples in Scaffold” on page 169](#), which describes how data is normalized in Scaffold.
- [“Quantitative Analysis Tests” on page 172](#), which describes the methods for inference available in Scaffold and how they are computed.

Label Free Quantitative Methods

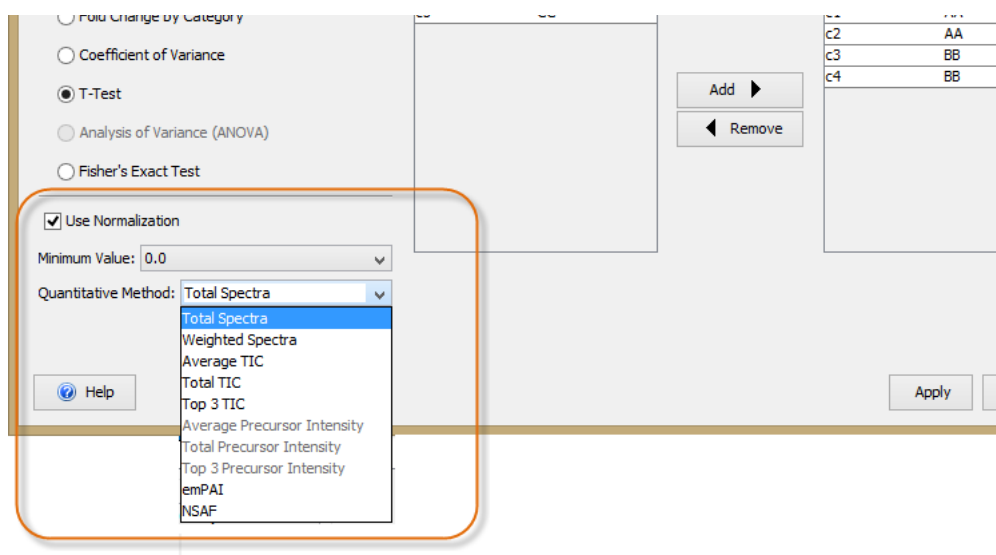
There are two widely used label free quantification strategies which are quite different in their approach and methods of accounting for the presence of proteins in a sample and a third one that is a sort of in between method.

- **Spectrum Counting**, which counts and compares the number of fragment spectra identifying peptides of a given protein
- **Precursor Ion Intensity**, which measures and compares the mass spectrometric signal intensity of peptide precursor ions belonging to a particular protein
- **Total Ion Count (TIC)**, which considers peak intensities from MS/MS spectra combined with counting of the spectra

For each of these main general methods Scaffold provides a number of variations that are commonly proposed in the standard literature:

- [Spectrum Counting](#)
 - [Total Spectra - \(default\)](#)
 - [Weighted Spectra](#)
 - [emPAI](#)
 - [NSAF](#)
- [Total Ion Count \(TIC\)](#)
 - [Average TIC](#)
 - [Total TIC](#)
 - [Top Three TIC](#)
- [Precursor intensity quantitation](#)
 - [Average Precursor Intensity](#)
 - [Total Precursor Intensity](#)
 - [Top Three Precursor Intensities](#)

Figure 9-1: Quantitative Method pull down list



How to select the proper quantitative method

When setting up an experiment researchers typically have a question in mind. The question determines the way the experiment is organized and conducted and also which quantitative method needs to be used to find answers to the question asked.

Typical questions asked in a mass spectrometry Proteomics experiment are:

1. Is anything changing?
2. How much is the amount of change I am dealing with?

Of the three main label free quantitative methods available in Scaffold, [Spectrum Counting](#) methods are the most reliable in answering question [number 1](#). The [Total Ion Count \(TIC\)](#) methods can answer both questions but not very well since they include limitations related to the counting of spectra while considering the peak intensities from MS/MS spectra.

[Precursor intensity quantitation](#) methods are very reliable in answering question [number 2](#).

Spectrum Counting

Scaffold includes the following Spectrum Counting methods:

Total Spectra - (default)

This method uses the sum of all the spectra associated with a specific protein within a sample which includes also those spectra that are shared with other proteins and is referred to as the Total Spectrum Count.

Weighted Spectra

This method uses the sum of all weighted spectra associated with a specific protein and within a sample, where the weight is a measure of how much a spectrum is shared by other proteins. For more details on how the weight is calculated see [Weighting Function](#).

emPAI

Spectrum Counting methods can also be used in the determination of absolute abundance of proteins. Initially the parameter used to measure this absolute abundance was the Protein Abundance Index (PAI) defined as the number of observed peptides divided by the number of all possible tryptic peptides from a particular protein, that are within the mass range of the employed mass spectrometer.

$$PAI = \frac{N_{observed}}{N_{observable}}$$

Where $N_{observed}$ is the number of experimentally observed peptides and $N_{observable}$ is the calculated number of observable peptides for each protein. In a subsequent refinement PAI was transformed into an exponential form called emPAI and defined as follows, see [Ishihama \(2005\)](#):

$$emPAI = 10^{PAI} - 1$$

NSAF

The [NSAF](#) quantitative method is useful when comparing the abundance of individual proteins in multiple independent samples and is typically applied to quantify the expression changes in various complexes. It is generally calculated using the number of spectra (SpC) identifying a protein divided by the protein length (L), referred to as Spectral Abundant Factor (SAF) and then normalized over the total sum of spectral counts/length in a given analysis. This means that SAF is then divided by the sum of SpC/L for all proteins in the experiment.

The NSAF values shown in [The Samples Table](#), when NSAF is selected as a quantitative value, are calculated using the NSAF strategy 2-a listed in Table 1 of [Zhang \(2010\)](#).

The calculation used in Scaffold translates to the following expression:

SAF = number of exclusive spectra/ length of proteins (expressed in number of amino acids)

The SAF value is then normalized using the regular Scaffold quantitative value normalization scheme, see [Normalization among samples in Scaffold](#), to derive the NSAF values shown in [The Samples Table](#).

NSAF calculations in Scaffold:

To check the calculation of NSAF in Scaffold the User should compute the SAF value for a

couple of proteins along the same column in the MS Sample view; to do so:

1. Select the Display option Exclusive Spectrum Count.
2. Select a protein from the protein list and annotate the exclusive spectrum count for that protein appearing in a specific MS sample.
3. In the Proteins View look for the number of amino acid in the protein.
4. Divide the exclusive spectrum count by the number of amino acids in the protein. This is the SAF for specific protein.
5. The values appearing in the Samples View when the Quantitative Value display option for NSAF is selected, is the normalized value of SAF. I
6. You can check the normalization factor for two values in same ms sample, it should be the same along a column.

Total Ion Count (TIC)

Scaffold includes the following [TIC- Total Ion Current](#) methods:

Average TIC

Average of all the TIC values of the spectra assigned to a protein. When selected the User needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Total TIC

Sum of all the TIC values of all spectra assigned to a protein. When selected the User needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Top Three TIC

Sum of the top three TIC values among the spectra assigned to a protein. When selected the User needs to adjust the **Minimum Value**: accordingly by selecting **Other...** in the pull down list.

Precursor intensity quantitation

Scaffold supports label-free quantitation based on precursor ion intensity for data from Proteome Discoverer, Mascot Distiller, and Spectrum Mill. Precursor intensity refers to the area under an MS1 spectrum peak corresponding to a specific peptide, whereas spectral counting counts the number of spectra identified for a given peptide. Scaffold provides three options for measuring a protein's precursor intensity in its Quantitative Methods drop-down menu in the [Quantitative Analysis...](#) dialog:

Average Precursor Intensity

This method takes the geometric mean of the peptide intensity values for a given protein.

When selected the User needs to adjust the Minimum Value accordingly by choosing **Other...** in the pull down list.

Total Precursor Intensity

The sum of all distinct intensity values for a protein. When selected the User needs to adjust the **Minimum Value:** accordingly by selecting **Other...** in the pull down list.

Top Three Precursor Intensities

The sum of the three highest peptide intensity values for a protein. If fewer than three peptides have intensity values, the intensities that are present are summed. When selected the User needs to adjust the **Minimum Value:** accordingly by selecting **Other...** in the pull down list.

Normalization among samples in Scaffold

To allow comparisons, Scaffold normalizes the MS/MS data. The User can then compare abundances of a protein between samples. The normalization scheme used works for the common experimental situation where individual proteins may be up-regulated or down-regulated, but the total amount of all proteins in each sample is about the same. It is not appropriate if the total amount of protein varies widely from one sample to the next.

In Scaffold there are two levels of summarization: the MS level which shows the samples run through the mass spectrometer; and the BioSample level, where BioSamples can contain one or more MS samples. Frequently the biological sample, or BioSample in Scaffold, is fractionated into multiple MS samples. Scaffold allows the User to view the MS samples within a BioSample or to combine all the MS samples into a single sample using the “MuDPIT” option. Normalization is performed at the MS sample level.

The normalization scheme in Scaffold adjusts the sum of the selected quantitative value for all proteins in the list within each MS sample to a common value: the average of the sums of all MS samples present in the experiment. This is achieved by applying a scaling factor for each sample to each protein or protein group adjusting in this way the selected value to a normalized “Quantitative Value”.

Note: For Precursor Intensities, since they operate at the peptide level, there might be various spectra that will show the same Intensity values. In the normalization scheme only one value will be considered for calculation purposes, for more information see [Precursor Intensity Quantitation in Scaffold](#) and [Performing Quantitation in Scaffold](#).

Note on Low abundance peptides

The normalization method used in Scaffold, as mentioned above, distorts the data if the total protein loaded varies considerably from sample to sample. This is due to the fact that low abundance peptides may be on the edge of detectability. If for example, sample A has a lot of protein loaded, the low abundance peptides may be detected. If sample B has much less protein loaded, these low abundance peptides might not be detected; that is, their spectral count is zero. No amount of scaling is going to change zero to any other number.

The User can view the normalized data selecting the **Quantitative Values** option from the Display Options pull down list in the Samples View. When viewing Quantitative Value (Normalized Total Spectra), the default quantitative method in Scaffold, if the User switches from this value to Total Spectrum Count and notices that all the values in one column change a lot compared to the values in the other columns, this is evidence of uneven protein loading. When this is happening it is important to be careful about how the data is used.

Normalization schemes not supported in Scaffold

There are more sophisticated normalization schemes that attempt to normalize the data in a way that allows the User to compare in a semi-quantitative way the abundance of one protein with another protein in the same sample. Scaffold does not support these schemes. This means that the User should exercise caution about trying to draw conclusions about the stoichiometry of the proteins from quantitative values as presented in Scaffold.

In particular, the User should be cautious about drawing conclusions about differential abundances for proteins where the spectral counts are small numbers. Scaffold tries to mitigate this problem by its treatment of [Missing values](#).

Missing values

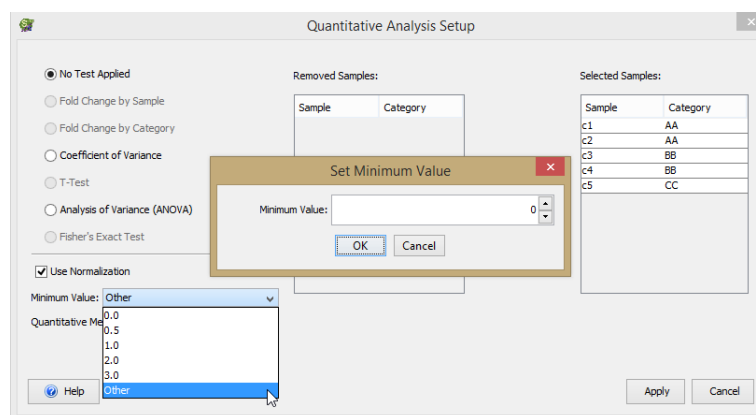
For differential protein expression tests, Scaffold replaces missing values with a specified [Minimum Value](#). Whenever a sample has no assigned spectra for a specific protein and that protein is found in a different sample, the specified minimum value is used to calculate the normalized values.

Minimum Value

The minimum value option allows the User to set a floor when calculating Label-Free quantitative values. Higher values will output shorter lists of highly confident changes; lower values will output longer lists that may contain less confident changes.

The minimum value is set by the User in the [Quantitative Analysis...](#) dialog through the Minimum Value pull down list and it defaults to 0. When selecting the option Other from the list the Set Minimum Value dialog opens allowing the User to record and use a minimum value different from the ones shown in the pull down.

Figure 9-2: Setting the minimum value



Whenever a sample has no assigned spectra for a specific protein and that protein is found in a different sample, the specified minimum value is used instead of zero for the sample with no assigned spectra. When Normalization is selected, the Missing Values are replaced by the set minimum value when quantitative values are calculated.

All quantitative values that are lower than the selected minimum value will also be replaced by the minimum value. This is true even if no statistical test is selected and the dialog controlling this value is grayed out.



-
- Scaffold does not display intensities lower than the minimum value.
 - The default minimum value is set at zero.
 - Select “Other” in the “Minimum Value” drop-down list to specify a custom value.
 - In the MS/MS Sample View parentheses indicate that the value shown in the cell was substituted with the minimum value.
 - In the BioSample View parentheses indicate that the subsumed value shown in the cell was derived from a set of values that contained values substituted with the minimum value.
 - For Fold change, if a zero appears in the denominator an INF is shown in the Fold Change column.

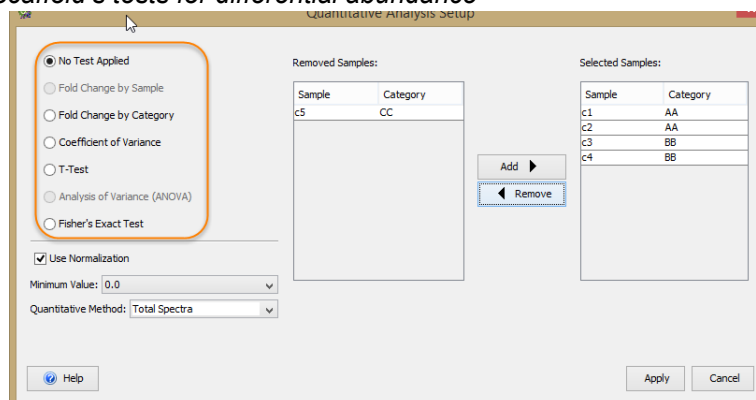
Quantitative Analysis Tests

Scaffold provides several tests to identify proteins which show different quantitative abundances in two or more categories. Which test to use depends upon the experimental design, particularly the number of replicates available.

The tests are available for selection through the [Quantitative Analysis...](#) dialog. In the dialog the User also chooses the categories being tested and the quantitative value being used.

The tests are based upon the data that is being displayed in the [The Samples Table](#). Adjusting the peptide and protein filters and thresholds or the ReqMod filter or toggling the [Show Lower Scoring Matches](#) changes the number of proteins shown in the table and the tests may select different proteins as having abundance level changes.

Figure 9-3: Scaffold's tests for differential abundance



Up to seven tests are potentially applicable, depending on the number of categories and replicate samples included in the categories:

- [Fold Change by Sample](#)

When a quantitative test is selected and applied, a column appears in [The Samples Table](#) with the test results. Sorting on this column brings the differentially expressed proteins together. These proteins can then be:

- checked for the quality of the spectra supporting the identification in the Proteins View
- checked for peptides shared between proteins in the Similarity View
- checked for differential expression using the Normalized Spectrum Counts Bar Chart in the Quantify View .

Fold Change by Sample

The simplest [Quantitative Analysis Tests](#) is the Fold Change, which reports by how much two variables differ. It is defined as the ratio of the quantitative value in one BioSample over the quantitative value in a second BioSample. The Fold Change by Sample can be used when

only two BioSamples are selected in the quantitative Analysis setup dialog. Because the specified Minimum Value replaces any Missing Values, if a zero appears in the denominator an INF will appear in the Fold Change column.

Notes:

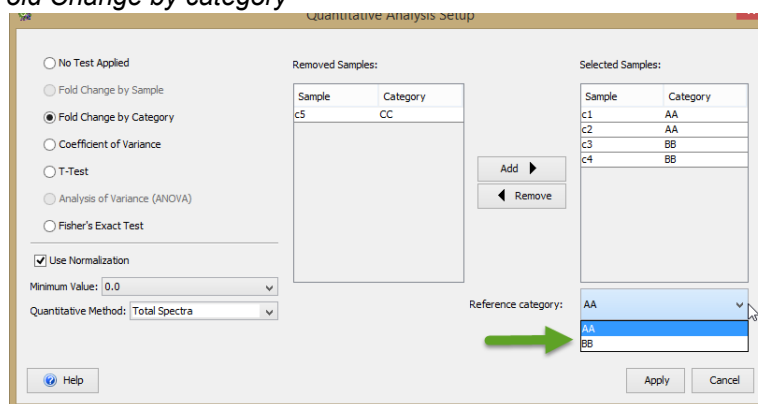
- Scaffold currently only shows the ratio, not the log base 2 of the ratio.
- Fold Change values need to be interpreted cautiously. A fold change of 2 is much more likely to be significant if the ratio is between 48 and 24 than if it is between 2 and 1. Scaffold's Q-Q scatter plot may help in this matter.
- If you sorting data based on the fold change, it is important to check both the top and the bottom of the sorted data. A 4 to 1 ratio will display as 4, but a 1 to 4 ratio will display as 0.25.
- If the fold change is less than 0.5 or 2.0, Scaffold colors the box green to help highlight possible proteins of interest. This does not necessarily signify any important statistical difference.
- Scaffold versions older than 3.5, if there are any missing values, report the ratio as 1.

Fold Change by Categories

The Fold Change by Categories is available for selection only when samples belonging to two different categories are selected in the Quantitative Analysis Setup dialog.

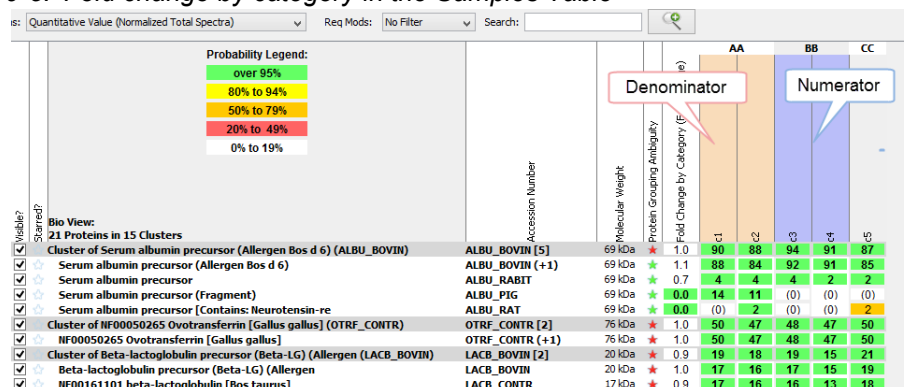
It is defined as the ratio between the average among the quantitative values of each BioSample included in one category versus the average of the quantitative values in the other category.

Figure 9-4: Fold Change by category



The Reference Category pull down list allows the selection of which category is used as denominator when calculating the Fold Change.

Figure 9-5: Fold change by category in the Samples Table



The column header of the samples included in the category considered as the numerator are highlighted in blue, the category considered as the denominator are highlighted in red.

Coefficient of Variance or Coefficient of Variation

A coefficient of variation or of variance (CV) can be calculated and interpreted in two different settings: analyzing a single variable or interpreting a model. The standard formulation of the CV, the ratio of the standard deviation to the mean, applies in the single variable setting. In the modeling setting, the CV is calculated as the ratio of the root mean squared error (RMSE) to the mean of the dependent variable. In both settings, the CV is often presented as the given ratio multiplied by 100.

The CV for a single variable aims to describe the dispersion of the variable in a way that does not depend on the variable's measurement unit. The higher the CV, the greater the dispersion in the variable.

The CV for a model aims to describe the model fit in terms of the relative sizes of the squared residuals and outcome values. The lower the CV, the smaller the residuals relative to the predicted value. This is suggestive of a good model fit.

$$C_v = \frac{\sigma}{|\mu|}$$

In Scaffold the User can select the Coefficient of Variance test only when he/she adds two or more BioSamples to the Select Samples Table in the Quantitative Analysis set up dialog.

- The CV is only defined for a nonzero mean. Because the CV is expressed as a percentage, Scaffold multiplies the ratio of the standard deviation to the mean by 100
- The CV is typically used to describe the dispersion of the variable independently of its measurement unit. The higher the CV, the greater the dispersion in the variable. For example, when analyzing four samples - A, B, C, and D - the coefficient of variance outputs how dispersed are the values in respect to their mean. A small coefficient of variance means that the four samples have values close together compared to their average value. If the coefficient of variance is big, then at least one of the four samples is

different, but it doesn't specify A, B, C or D. Examining Scaffold's Quantitative Value Bar Graph helps determine which it is.

- Coefficient of Variance is typically used in place of an ANOVA test when not enough replicates are available to give sufficient statistical power to apply ANOVA.

T-Test

The T-test is a measure of the distance between the mean of the replicate samples in one category from the mean of the replicate samples in another category. This distance is scaled by the standard deviation of the replicates. The results of a T-test is reported as the probability (p-value) that this distance between means could occur by chance.

To be able to apply the T-test in Scaffold, the BioSamples in the experiment need to be organized at least in two different categories, see [Organize Samples In Categories](#). Among the various samples in the experiment only samples belonging to two different categories need to be included in the Selected Sample table in the Quantitative analysis set up dialog to have access to the T-test option.

Each of the two categories should include three or more replicate BioSamples. Examples of typical categories are “treated/untreated”, “disease/control” or “cell line1/cell line2”. Since the test is computed using quantitative values most of the time normalized, the User should keep in mind potential issues surrounding [Missing values](#) and Normalization as described in [Normalization among samples in Scaffold](#).

A small p-value means that the BioSamples in one category are most likely different from those in the other category. A threshold or alpha level or significance level of 0.05 is commonly used to assess how statistically significant the result of the T-test is. This value should be appropriately adjusted in Proteomics experiments since differences are evaluated among many proteins at once, see

The T-test is generally considered a fairly robust test. This means that even if its basic assumptions are violated somewhat, it still tends to be fairly reliable at separating the categories which are the same from those that differ. Some researchers believe that spectral data should be transformed in some way, for example by taking its log, before doing a T-test. Other researchers may think that 3 replicates is not enough to apply the T-test. Still others believe that more advanced non-parametric tests would work better. So if the T-test gives a borderline result, the User may want to check it carefully. But if the T-test has a very small p-value, the robustness of the T-test means that it is unlikely that a more sophisticated statistical analysis will give a different result.

If the User tries to push things by computing the T-test with less than 3 replicates, it is unlikely to give informative and trustworthy results. The [Fisher's Exact Test](#) may be more appropriate for samples with few replicates and low abundance proteins with few spectral counts.

Sometimes people make a distinction between technical replicates and biological replicates. This is a more advanced statistical analysis concept and is not supported by Scaffold.

ANOVA

The ANOVA (Analysis of Variance) is an analysis method for testing equality of means across treatment groups or categories. It tells if there are differences among categories. The result of the test is a p-value which when low indicates a large probability for variation among the different categories considered for the test.

The ANOVA test in Scaffold requires three or more replicates in the categories. Like the T-test, having fewer than 3 replicates is untrustworthy and more replicates are better.

Like the Coefficient of Variance test, the ANOVA test shows that something is different, but it doesn't tell what categories are different from each other. Checking the Bar Chart in [The Quantitative Value pane](#) helps understanding which category is different.

The ANOVA test in Scaffold is a simple one-way ANOVA. More sophisticated ANOVA tests are beyond Scaffold's capability.

Before applying the ANOVA test the User should understand the issues regarding [Protein Grouping Ambiguity](#), [Missing values](#), and Normalization as described in [Normalization among samples in Scaffold](#).

Fisher's Exact Test

The **Fisher's Exact Test, like the T-test**, compares the relative abundance between two sample categories. It is used in the analysis of contingency tables where sample sizes are small. It is called exact since it calculates the significance of the deviation from the null hypothesis with an exact method not using an approximation dependent on the size of the sample statistics. This means that the Fisher's exact test is more appropriate than the T-test if there are fewer replicates. Like the T-test, the Fisher's Exact Test produces a p-value.

Scaffold calculates the Fisher's exact test p-value according to a model discussed in [Zhang \(2006\)](#). The paper performs a systematic analysis of the various approaches to quantify differential expressions among different experiments, Particularly it describes how and when it is reasonable to apply a Fisher's Exact test in a pair wise experiment.

As described in the paper, to calculate the Fisher exact test for a target protein, Scaffold arranges the spectral counts for a pair of categories into a two-way contingency table where the first row contains the counts for the target protein in each category and the second row contains the rest of the counts for the rest of proteins listed in the Samples table. The test is based on the assumption that the row and column totals are fixed, which means that any entry in the table completely determines the others. The probability assigned to a particular arrangement of spectral counts in the table is calculated using a hyper-geometric distribution.

The p-value assigned by the Fisher's Exact test to the target protein is the sum of all p-values over all the possible configurations of the two-way table that have p-values less or equal to the initial target protein two way table (which means integrating over the tail of the distribution).

Multiple tests

The statistical quantitative tests available in Scaffold give some measure of how different the BioSamples in the various categories are. These measures are either a ratio, a coefficient of variance, or a p-value.

How big do these measures have to be before they are significant? The naive approach is to set an arbitrary value, say a 2-fold change, or a p-value of less than a significance level of 0.05. This is what Scaffold does when it colors the fold change or p-value green. This probably includes some proteins that should not be labeled as differentially expressed but have gotten on the list by chance.

A better approach is to sort the proteins by their p-value so the small p-values are on top. The proteins at the top of the list are the most likely to be accurately classified as differentially expressed. As the User goes down the list the confidence in the classification should diminish.

The question becomes where the User should draw the threshold line. Scaffold leaves this to the judgment of the researcher or User. The issue is complicated because considering a set of statistical inferences all together increases the chance of falsely identifying a difference as significant, type I error. The User may want to seek guidance from a statistician and ask about methods of correction for multiple testing in analyzing the T-test and ANOVA results.

Chapter 9

Quantitative Methods and tests

Chapter 10

Precursor Intensity Quantitation

An increasingly popular option for quantitative Proteomics is Precursor Intensity Quantitation, which offers a good compromise between the accuracy of labeled techniques and the simplicity and lower cost of label-free quantitation. This method relies on measuring the signal intensity of the peptide precursors representing a specific protein at the MS level and comparing these intensities across samples. Both Scaffold and Scaffold Q+/Q+S support this method in different ways.

Precursor Intensity Quantitation in Scaffold

Scaffold is designed to provide easy and confident validation, visualization and quantitation of search results. It does not read raw files and does not have direct access to precursor information; instead it reads intensity data already computed by the identification software. Currently, Scaffold is able to obtain precursor intensity information from Thermo Proteome Discoverer, Mascot Distiller, Agilent Spectrum Mill, and MaxQuant files. Scaffold normalizes precursor intensity values across samples and calculates fold changes at the BioSample or Category level. Statistical tests of differences in the calculated intensities are also offered, including the T-Test, ANOVA and Coefficient of Variance as appropriate to the experimental design.

Because of its dependence on search results, Scaffold's approach to Precursor Intensity Quantitation is to work backward from the peptides that have been identified through their MS/MS spectra and compare the intensities of the MS peaks from which they were derived. By contrast, some programs align the MS peaks of all samples and calculate their intensities. The peaks that appear to be biologically important are subsequently identified. This method has the advantage of providing quantitative information for low abundance proteins, but it relies on complicated peak alignment algorithms and may in some cases incorrectly identify the corresponding peaks.

Working from identifications is simpler, since it obviates the need for retention time warping and peak alignment, and it has the advantage of depending on more reliable data. On the other hand, in this approach missing values become an issue. Often a peptide is identified in one sample and not in another, producing a missing value even if there may have been a detectable MS peak in the corresponding position in the second sample. Generally, however, higher abundance peptides are more likely to be identified and MS peaks that do not result in identified spectra are relatively weak signals, minimizing the effect of treating them as missing values. Scaffold further reduces the effect by choosing algorithms that are less sensitive to missing values: for example it uses the geometric mean rather than the average in calculating protein level fold-changes.

Calculation of Precursor Intensities

Precursor Intensity Quantitation is based on the principle that the area of the peak in the MS1 chromatogram provides a measure of the relative abundance of the corresponding peptide in the sample. Peptides are identified based on their MS/MS spectra, and then the corresponding MS1 peaks are identified in each LC-MS/MS run. The areas under these peaks are calculated and normalized and their ratios are used as a measure of the relative abundance of the peptides in different samples. Relative quantities of proteins are estimated by combining the precursor intensities of the constituent peptides in various ways.

The following illustration of the typical LC-MS/MS analysis of a peptide is reproduced from [Lai \(2013\)](#).

Figure 10-1: Identification of a peptide through LC-MS/MS analysis

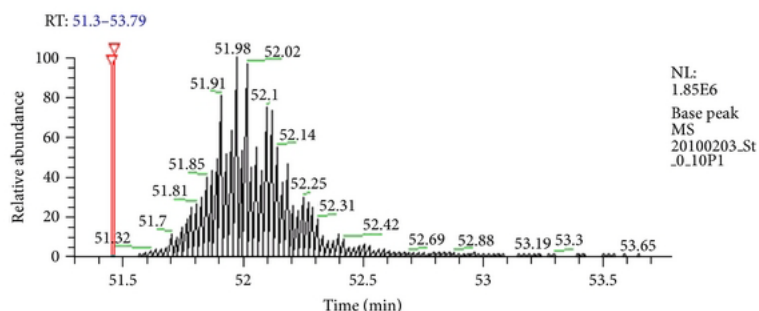


Figure 10-1-a: The peptide is eluted from the LC column and its ion intensity is plotted as a function of the retention time.

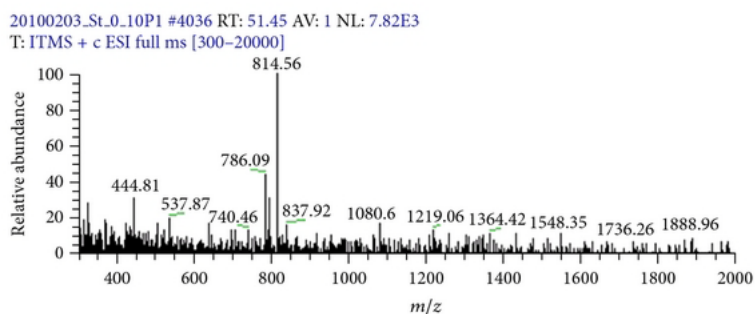


Figure 10-1-b: At the first scan time shown in red in (a), a full MS scan is performed. The ion with m/z 786.09 is selected as a precursor ion for MS/MS analysis.

20100203_St.0_10P1 #4037 RT: 51.46 AV: 1 NL: 1.31E2
T: ITMS + c ESI d w full ms2 786.09@cid35 [205-20000]

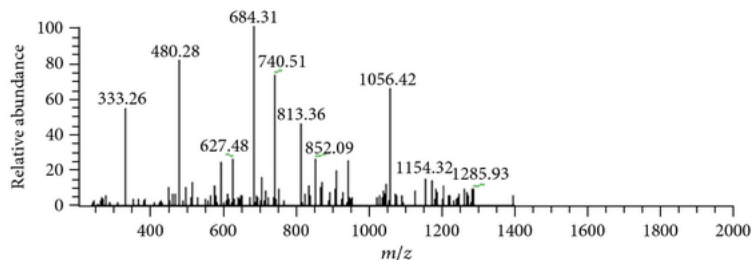


Figure 10-1-c: At the next scan (also shown in red in (a)) , an MS/MS scan is performed, providing peptide fragmentation information for peptide identification.

Once a peptide has been identified, a program can go back to the MS1 scans and find a series of spectra which contain peaks corresponding to the same peptide as it continues to elute from the column. These spectra are then aligned and the intensities of the peaks for the specific m/z value which represents the parent ion of this peptide are plotted against the retention time, giving an extracted ion chromatogram.

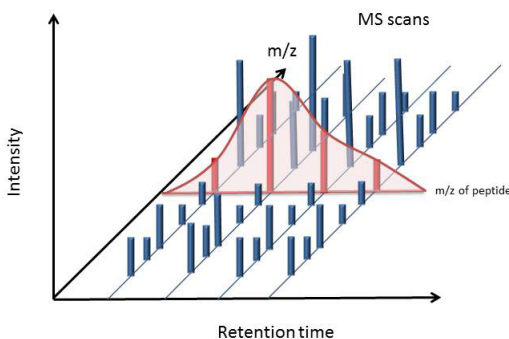


Figure 10-2: The intensities of the MS peaks at the same m/z value are plotted as a function of the retention time. The area under this curve (enclosed in red) is the precursor ion intensity.

In the extracted ion chromatogram, a curve is fit to the intensities at a specific m/z . The area under this curve represents the total amount of the specified peptide that eluted. Scaffold reads these values from its input files and uses them to do quantitative analysis.

Preparing Data for Precursor Intensity Quantitation in Scaffold

Scaffold reads precursor intensity information from various identification programs provided that the User has requested this type of quantitation during the search. Following are

instructions for preparing input files for quantitation in Scaffold:

Proteome Discoverer

Proteome Discoverer provides a workflow template for computing precursor intensity values. The template **WF_LTQ_Orbitrap_Sequest_Precursor_ions_Area_Detector** can be used as a starting point, and the search engine choice or instrument settings may be changed. Scaffold reads the precursor intensities from the MSF file.

Mascot Distiller

When setting up the Mascot search, select **Average[MD]** as the quantitation method. When the search is complete, in Distiller select **Analysis>Calculate XIC**, and then **Analysis>Quantitate**. Export the results as an XML file using **Analysis>Quantitative Report>Save as XML**. Also create an ROV file by saving the project with **File>Save Project As...** Place the ROV file and the XML file in the same directory, and if the DAT file is not accessible directly from the Mascot Server, also place that file in the same location. **Load only the XML file into Scaffold.**

Spectrum Mill

No special settings are required. Load the entire Spectrum Mill results directory into Scaffold.

MaxQuant

MaxQuant 1.3 will only compute precursor intensity when two or more raw files are processed together. Each of the samples to be compared must be labeled with a different experiment name in the experiment.txt file.

Generally, all MaxQuant results in a single directory load into Scaffold as a single sample. For precursor quantitation, however, the samples to be compared must be loaded into different BioSamples. Accordingly, Scaffold has a special dialog that opens when the program recognizes the presence of an experiment file. To place each experiment into its own BioSample, from the loading wizard select the MaxQuant output directory and click **Add to Import Queue** then when the dialog appears, select the first experiment. Click Next, then **Add another BioSample** and select the same directory, but choose a different experiment from the dialog box.

In **MaxQuant 1.4**, precursor intensity may be computed even when analyzing a single raw file if the user selects the Label Free Quantitation option. Individual results may then be loaded into separate BioSamples in the usual way and used for Precursor Intensity Quantitation in either Scaffold or Scaffold Q+.

If two or more raw files are analyzed together in MaxQuant 1.4 with the Label Free Quantitation option selected, and no **Experiment.txt** file is provided, they form a single combined folder which loads into Scaffold as a single sample. In this case, Scaffold and Scaffold Q+ are unable to perform Precursor Intensity Quantitation. It is possible, although not required, in MaxQuant 1.4 to create an experiment file. The experiments can be named through the MaxQuant 1.4 GUI, and then an experiment file can be exported by right-

clicking and choosing **Export**. The user should name the file **Experiment.txt** and then Scaffold will recognize it and loading can proceed as for MaxQuant 1.3 results.

Performing Quantitation in Scaffold

Scaffold reads the precursor intensity values already computed by the search engine software from the input files. For each peptide-spectrum match, it reports the intensity value in the Peptides Table in the upper right of the Proteins View ([Figure 10-3](#)).

Charge	Delta ...	Delta ...	Rete...	Intensity	TIC	Start	Stop	# Ot...	Other Pr
2	-0.011	-6.5	2720	1.88E7	728700	2	15	0	
2	-0.011	-6.5	2780	1.88E7	190600	2	15	0	
3	-0.0086	-5.2	2780	8.33E7	191400	2	15	0	
3	-0.0091	-5.5	2720	8.33E7	559200	2	15	0	
3	-0.0096	-5.8	2880	403000	63620	2	15	0	
3	-0.0095	-5.9	1700	20800	301700	2	15	0	
2	-0.018	-7.9	5230	30000	547400	16	33	0	
2	-0.013	-6.0	3990	195000	2466000	16	33	0	
4	0.0063	2.8	3990		28220	16	33	0	
3	-0.013	-6.0	5250	5920000	256800	16	33	0	
4	0.0048	2.1	5250		24070	16	33	0	
3	-0.015	-6.6	5190	5920000	18650	16	33	0	
3	-0.015	-6.9	3990	4260000	522900	16	33	0	
4	-0.013	-5.7	5190	5530000	10210	16	33	0	
3	-0.012	-6.0	5600	2.33E7	519400	18	33	0	
3	-0.013	-6.4	5540	2.33E7	93110	18	33	0	

Figure 10-3: Precursor Intensities in the Peptides Table

Scaffold provides three methods of using these values to perform relative quantitation at the protein level. These methods are available through **Experiment > Quantitative Analysis** or by clicking on the bar-graph icon at the top of the screen. Either of these methods brings up the **Quantitative Analysis Setup** dialog (Figure 6). Because it is a relative quantitative method, when using label-free quantitation, it is necessary to select at least two samples. It is also important to adjust the **Minimum Value** setting to a value that is appropriate for intensities. Values other than zero require the use of the **Other** option in the dropdown. A checkbox allows the user to choose whether or not to normalize between samples.

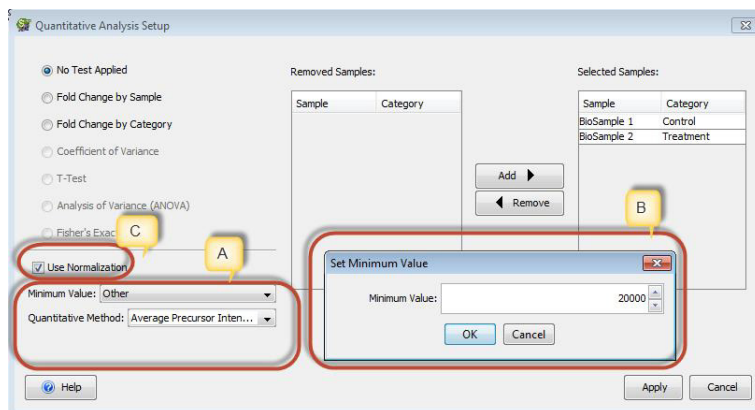


Figure 10-4: The Quantitative Analysis Setup Dialog – A. Selecting the Quantitative Value, B. Specifying the Minimum Value, C. Selecting Normalization option

There are a number of options for combining peptide precursor intensities to provide an estimate of relative quantities at the protein level. Scaffold provides three methods: **Average Precursor Intensity**, **Total Precursor Intensity** and **Top 3 Precursor Intensity**. These are calculated from the **Intensity** values shown in the Peptides Table as follows:

Valid	...	Sequence	Modifications	Charge	Intensity	Observed
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)	Oxidation (+16), Oxidation (+16)	4	2.61E7	131
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)	Oxidation (+16), Oxidation (+16)	3	2.61E7	171
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(K)	Oxidation (+16), Oxidation (+16)	3	2650000	332
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Oxidation (+16)	3	5610000	434
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Oxidation (+16)	2	6100000	650
<input checked="" type="checkbox"/>	1.0	(K)AGIFQSAK(-)		2	5.13E7	411
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)	Oxidation (+16)	2	1390000	854
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)	Oxidation (+16)	3	6060000	570
<input checked="" type="checkbox"/>	1.0	(K)KAYAEFYR(N)		2	2430000	524
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)	Deamidated (+1), Oxidation (+16)	2		855
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)	Oxidation (+16)	3		570
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(K)	Oxidation (+16)	3	980000	521
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(K)	Oxidation (+16)	2	3000000	790
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42), Oxidation (+16)	2	6.65E7	671
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42), Oxidation (+16)	2	6.65E7	671
<input checked="" type="checkbox"/>	1.0	(R)NYDSMKDFEEMRK(A)		3	840000	846
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42)	2	3.81E7	663
<input checked="" type="checkbox"/>	1.0	(M)SSGALLPKPQMR(G)	Acetyl (+42)	2	3.81E7	663

Figure 10-5: Intensity values for a singleBioSample for a specific protein - A. Note that often multiple MS2 spectra are collected from a single MS1 spectrum. This results in duplicate reports of the same Intensity value. Scaffold counts each value only once. B. Intensities for different charge states of the same peptide are summed to give the total intensity for that peptide.

First, peptide intensity values are calculated. As shown in Figure 10-5, duplicate intensity values for the same peptide are discarded. If there are multiple peptide-spectrum matches with the same peptide sequence and modifications but with different intensity values, their intensities are summed and the sum is used as the intensity value for that peptide. The peptide intensity values are then used in the following calculations:

- **Average Precursor Intensity:** The geometric mean of the peptide intensity values for a given protein.
- **Total Precursor Intensity:** The sum of all distinct intensity values for a protein.
- **Top 3 Peptides Precursor Intensity:** The sum of the three highest peptide intensity values for a protein. If fewer than three peptides have intensity values, the intensities that are present are summed.

When one of these methods is selected through the **Quantitative Method** drop-down, it becomes available for display in the Samples View. Choosing **Quantitative Value** from the **Display Options** drop-down causes the Samples View to show precursor intensity values calculated according to the selected method in the Samples Table. The name of the method is displayed in Display Options and if the values have been normalized, that is also indicated (Figure 10-6).

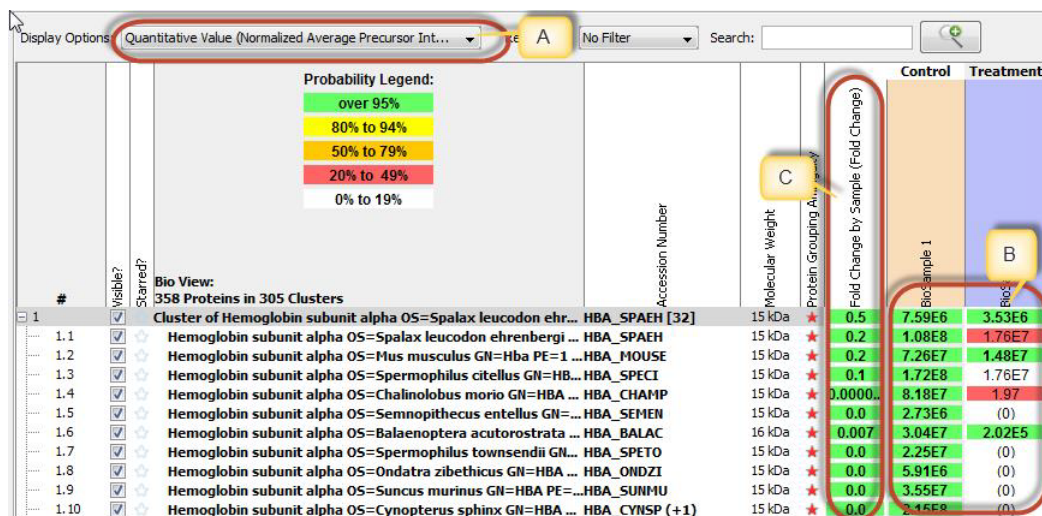


Figure 10-6: The Samples View – A. Quantitative Method selected B. Protein-level intensity values C. Fold Change calculated from the precursor intensity values.

Using the Quantitative Values based on precursor intensity, Scaffold can also calculate fold change at either the BioSample or the Category level. The desired fold change option is specified in the Quantitative Analysis Setup, which also allows selection of which BioSample or Category should serve as the reference (Figure 10-7).

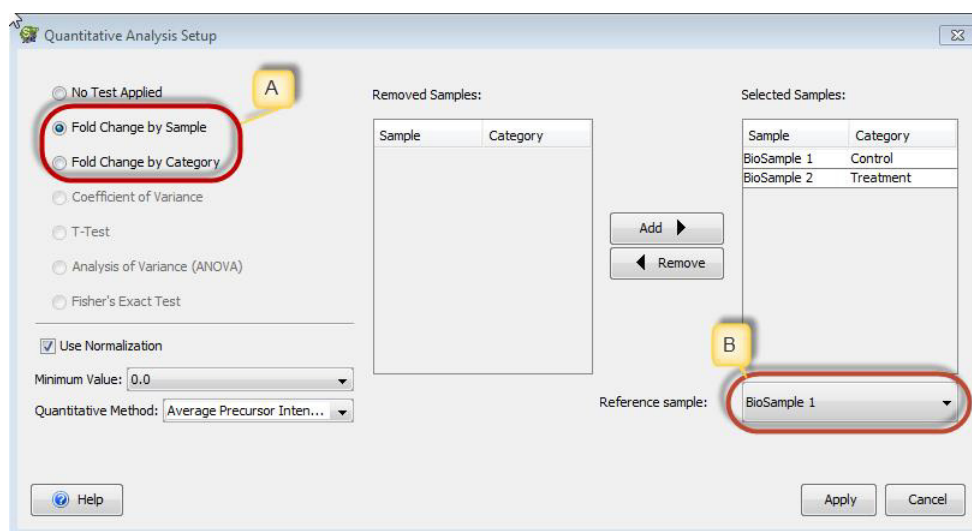


Figure 10-7: Requesting display of Fold Change – A. Choice of Fold Change by Sample or by Category, B. Specification of Reference Sample or Category.

When a Fold Change option is selected, an additional column is displayed in the Samples View. Fold Change is based on the Quantitative Method selected in the Quantitative Analysis Setup dialog even if a different display type (such as Total Spectrum Count) is displayed in the Samples View.

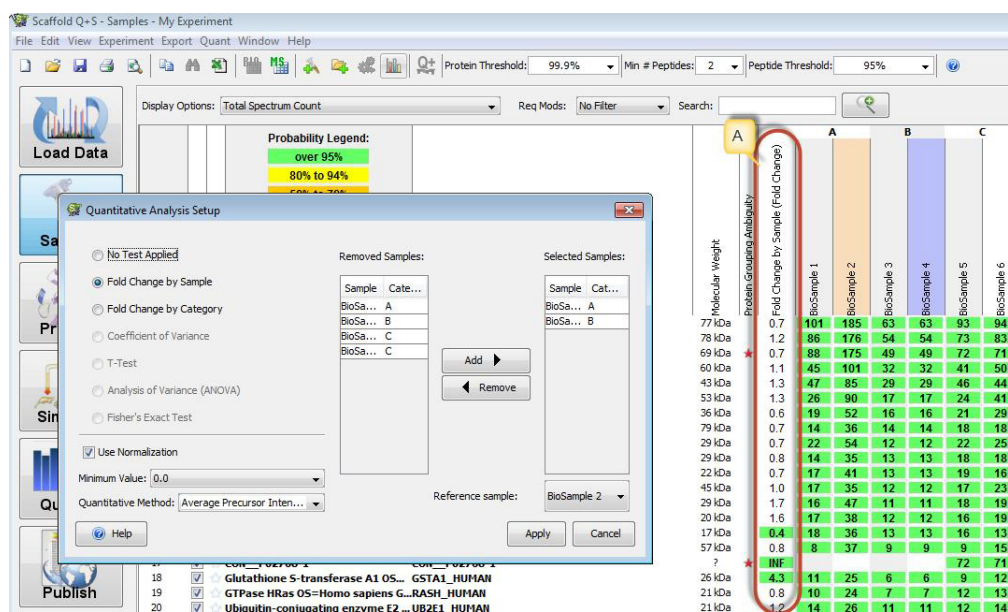


Figure 10-8: Fold Change by BioSample – A. The Fold Change column, showing the ratio of the Average Precursor Intensity of BioSample 4 to the Average Precursor Intensity of BioSample 2.

Fold Change by Sample is only available if exactly two BioSamples are selected for quantitation. It displays the ratio of the quantitative value of the non-reference BioSample to the quantitative value of the reference BioSample for each protein. The reference Sample is indicated by peach coloring in the column header, and the sample being compared is indicated by a purple header.

If samples from exactly two categories are selected, Fold Change by Category is available for display.

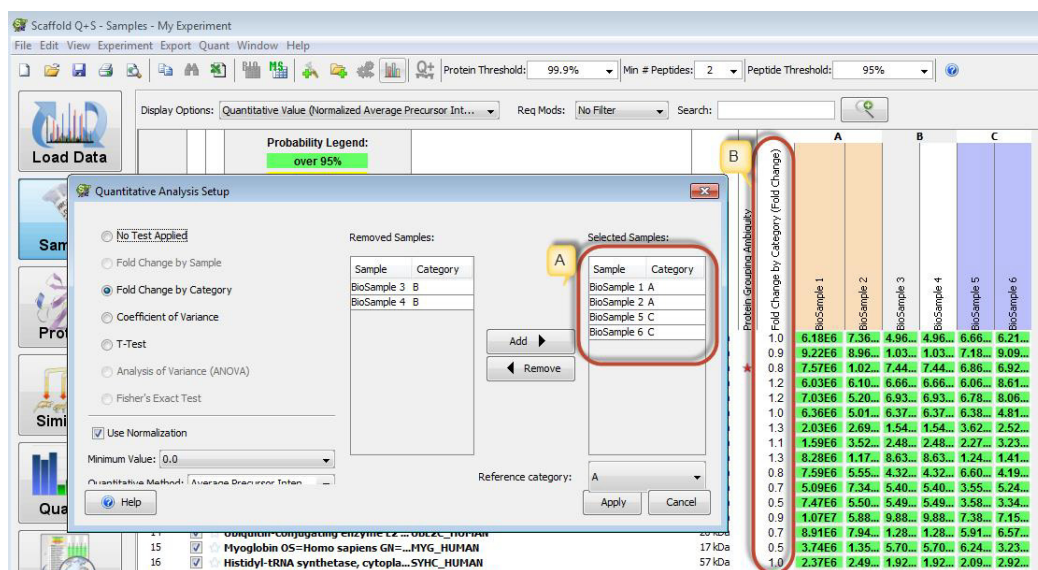


Figure 10-9: Fold Change by Category – A. Selected samples must belong to exactly two categories. B. The Fold Change values are displayed in a column in the Samples View.

The fold change values represent the ratio of the average of the quantitative values of the selected samples in the comparison category to the average of the quantitative values of the samples in the reference category.

Chapter 11

Reports

A variety of reports are available in Scaffold to assist the User in interpreting and working with quantitative analysis data. All the reports are available from the Export option on the Scaffold main menu. Every report is saved in a predefined format, and in the same directory as its quantitative analysis data.

The User cannot change the report format, but can always select a different location in which to save the report. When the User saves a Scaffold ProtXML report, he/she must provide a name for the report. When the User saves an Excel report, a default name in the format <Report Name><Scaffold File name> is provided for the report, but their values can always be changed. Finally, the User can open and view any Excel report such as the Publication report in Excel or another spreadsheet application, but the User might need to specify that the report file is a tab-delimited file to do so.

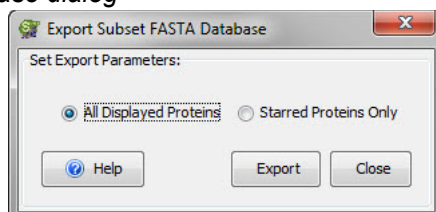
The following reports are available in Scaffold:

• Subset Database	190
• Spectra	190
• ProtXML report	191
• mzIdentML	191
• ScaffoldBatch... ..	196
• ScaffoldBatch Archive... ..	196
• Exports compatible with Excel	197

Subset Database

The command **Export > Subset Database** exports a FASTA database subset of the original sequence database used for searching the imported data.

Figure 11-1: Subset Database dialog



When selected the command opens a dialog containing the following export parameters:

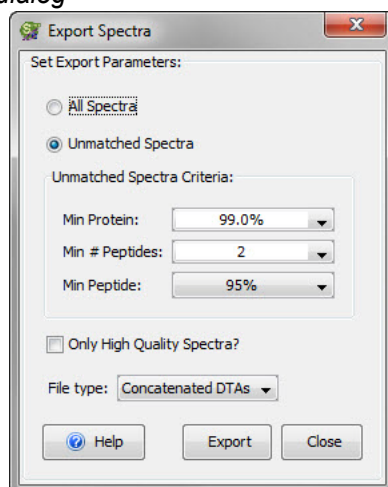
- **All Displayed Proteins** - when chosen the created subset database contains only proteins appearing in the proteins list. Adjusting the protein and peptide thresholds determines which identified proteins are included in the list and subsequently in the subset database. (More restrictive parameters result in fewer proteins in the exported database.)
- **Starred Proteins Only** - When selected the subset database will include only proteins labeled using stars.

The exported subset can facilitate a more thorough search for protein modifications with the original search engine.

Spectra

The menu command **Export > Spectra** exports spectra loaded in the current Scaffold experiment as peak lists. A list of different formats is available for the User to choose how to save the exported peak list.

Figure 11-2: Export Spectra dialog



When selected the command opens a dialog containing the following parameters options:

- **All Spectra** - This option exports all the spectra loaded in the current Scaffold experiment.
- **Unmatched spectra** - This option exports only spectra that do not meet the filters criteria set by the User to allow further targeted searches on these types of spectra. The criteria are based on the probabilities assigned by Scaffold through its scoring algorithms.

Unmatched Spectra Criteria:

- Min Protein
- Min # Peptides
- Min Peptide
- **Only High Quality Spectra?** - When selected Scaffold chooses for export only those spectra that identify peptides with probabilities higher than 50% or if the peptide probability happens to be lower, the spectra has to be assigned to proteins that have a probability of at least 95%.
- **Types of peak list files:**
 - Concatenated DTAs
 - Individual DTAs
 - Mascot MFGs
 - Micromass PKLs
 - SEQUEST MS2s

ProtXML report

Exports all quantitative data in the protXML format, which is an open XML file format for the storage of data at the raw spectral data, peptide, and protein levels. This format enables uniform analysis and exchange of MS/MS data generated from a variety of different instruments, and assigned peptides using a variety of different database search programs.



See Molecular Systems Biology, 1:2005.0017 for more information.

mzIdentML

Scaffold fully supports the mzIdentML standard format for Proteomics data developed by the HUPO Proteomics Standards Initiative (Proteomics Informatics Standards group). A description of the standard specifications is available at www.psidev.info/mzidentml and a Java desktop software for validating mzIdentML can be downloaded at code.google.com/p/psi-pi/downloads/list.



Scaffold supports both mzIdentML 1.0.0 and the latest version 1.1.0.

Exports are compatible with:

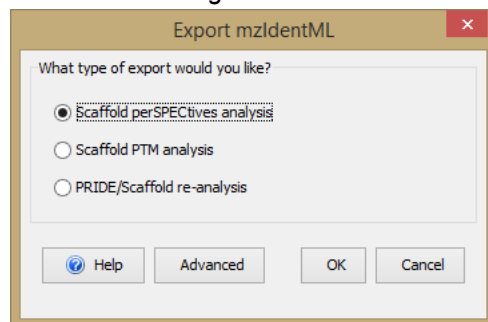
- The PRIDE-XML converter.
- SKYLINE for building spectral libraries. See [Creating a spectral library in Skyline](#)

Selecting the menu command **Export > mzIdentML** opens the **Export mzIdentML** dialog where the User can easily customize his/her mzIdentML exports.

The **Export mzIdentML** dialog shows three basic options for creating mxIdentML exports optimized for the following uses:

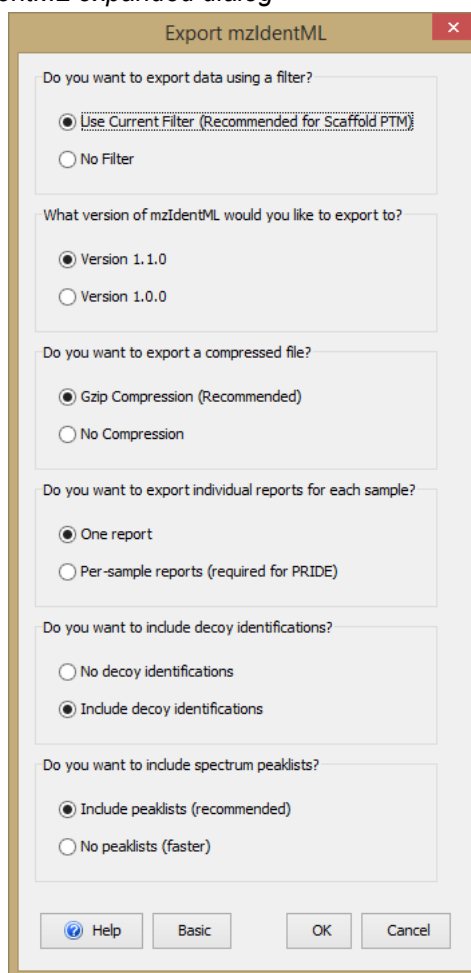
- **Scaffold perSPECTives analysis**
- **Scaffold PTM analysis**
- **PRIDE/Scaffold re-analysis** -- This export is suggested both for loading data in PRIDE or reloading data in Scaffold using mzIdentML instead of the regular search engines files.

Figure 11-3: Export mzIdentML short dialog



Clicking the **Advanced** button expands the dialog to show the full list of options available to further customize the mzIdentML export.

Figure 11-4: Export mzIdentML expanded dialog



The available options are the following:

- Selection of the list of proteins to include in the MzIdentML through the set filters
- Selection of the version of the file exported - Scaffold supports the latest version of MzIdentML the 1.1.0 and previous ones.
- Selection of the type of compression
- Selection of the number of reports exported- With multiple BioSamples it is possible to create mzIdentML exports for each BioSample included in the experiment
- Inclusion of decoys
- Inclusion of peak lists - The peak list is saved using the MGF format.



*The mzIdentML export creates one or more *.MZID files and a series of *.MGF files, if the inclusion of peaks option is selected, saved in a newly created meaningfully named directory.*

Creating a spectral library in Skyline

Skyline is a popular application used to create and iteratively refine targeted methods for proteomics studies, see [Skyline](#). It also provides tools to build spectral libraries from validated peptide spectrum matches.

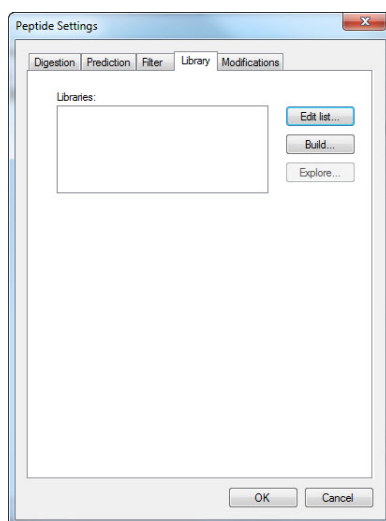
Scaffold experiments can include a variety of validated spectrum matches coming from different sources and analyzed using multiple search engines. This means that Scaffold experiments can be a good source for validated spectrum matches. The way Scaffold exports spectral identification results is through mzIdentML reports and associated peak lists in MGF format. Within these files Scaffold embeds precursor intensity, retention time, and a reference to the original RAW file, which are requirements for creating transition libraries in Skyline. The mzIdentML exports are now compatible with Skyline and can be used to create spectral libraries within that application.

Furthermore, Scaffold supports a large variety of search engine reports, some of which are not currently compatible with Skyline. In particular Proteome Discoverer is a common platform for MS/MS based proteomics which is now compatible with Skyline using Scaffold as an interface.

The User can create a spectral library in Skyline following these instructions:

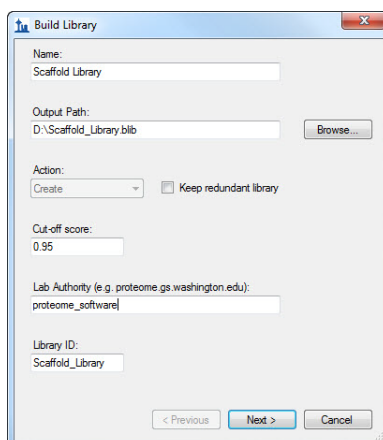
1. From a Scaffold experiment select the menu option **Export > mzIdentML...**, the Export [mzIdentML](#) dialog opens. In the dialog select the option Scaffold perSPECTives analysis and then click **Advanced** see [Figure 11-4](#).
2. To the question: **Do you want to export a compressed file?** select the answer **No compression**. Click **OK** to save the mzIdentML files.
3. To create a spectral library in Skyline go to the menu option **Settings > Peptide Settings....** The **Peptide Settings** dialog opens onto the tab **Library**.

Figure 11-5: Skyline: Peptide Settings, Library tab



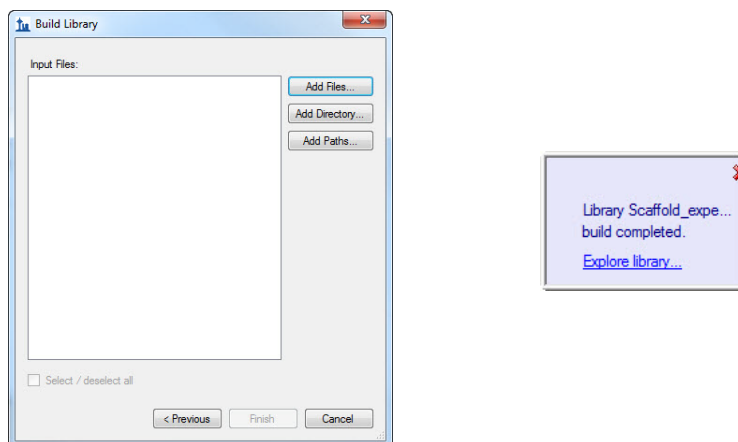
4. Click **Build** to add a new library. The Build Library wizard opens.

Figure 11-6: Skyline: Build library wizard initial page



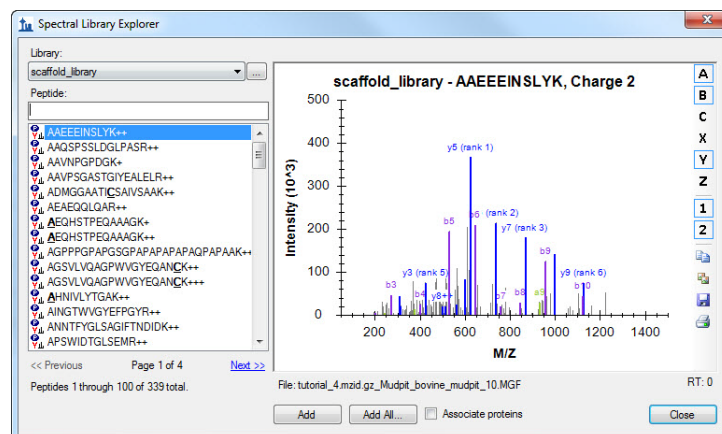
5. Assign a name to the library and if needed adjust the output path where the library is saved, click **Next**.
6. In the new page of this wizard click **Add Files** and point Skyline to the location where the Scaffold *.MZID is saved.
7. Once the file is selected it appears listed in the Input file text area. Click **Finish**.

Figure 11-7: Skyline: wizard and message when the library is loaded



8. A message appears once the library is built up and ready for use.
9. The newly built library is listed in the text area Libraries in the **Peptide Settings** dialog, **Library** tab.
10. To view the newly loaded library first select it in the **Peptide Settings** dialog and then choose **View > Spectral Libraries**. The Spectral Library Explorer dialog opens.

Figure 11-8: Skyline: Spectral Library Explorer



ScaffoldBatch...

ScaffoldBatch is a command line version of Scaffold designed to run in the background of a system where it is installed. ScaffoldBatch reads its commands from an XML type driver file (*.SCAFML) rather than from the graphical user interface (GUI) and creates an *.SF3 Scaffold experiment.

When selected the command option **Export > ScaffoldBatch** creates a SCAFML driver file that if run in ScaffoldBatch reproduces the Scaffold experiment from which it is exported.

SCAFML files are often used as an interface between Scaffold and another program. A number of labs have created custom software that uses SCAFML driver files as an interface between a Laboratory Information Lab System (LIMS) and Scaffold.

Commercial versions of ScaffoldBatch are available bundled into Sage N Research's Sorcerer and Genomics' Proteus Analytics. Matrix Science has an interface between their Integra LIMS system and Scaffold that uses SCAFML files.

ScaffoldBatch Archive...

When selected the command option **Export > ScaffoldBatch Archive** bundles into one package a SCAFML driver file that, if run in ScaffoldBatch, reproduces the Scaffold experiment from which it is exported, along with all the files that are referenced in the *.SCAFML driver. This package typically contains the input files and the FASTA database and it is saved in a compressed format like zip or tar. Before running it on a computer that

has ScaffoldBatch installed the User needs to unzip or untar the package.

Exports compatible with Excel

Scaffold provides a number of tab delimited reports containing different types of information related to the analysis performed in the current Scaffold experiment:

- [Publication report](#)
- [Samples report](#)
- [Spectrum report](#)
- [Peptide report](#)
- [Protein report](#)
- [Current View report](#)
- [Complete report](#)

How to open Scaffold reports in Excel

The exported reports can be viewed in Microsoft Excel for further analysis of the data they contain. When importing any of these reports into the spreadsheet, it may be necessary to specify that the report is TAB delimited. Excel may also show its Text Import Wizard the first time the User opens an exported Scaffold report. Selecting delimited file, then tab-delimiters completes the conversion to the Excel format. Saving as an XLS file avoids repeating the conversion in the future



To create an export that includes the GO annotations see [Samples report](#)

Publication report

The Publication report lists the data analysis information required for publication in a number of the Proteomics journals. This report is a copy of the information reported on the Publish view.

The top of the report lists in a structured manner:

- How the peak lists were generated
- What databases were searched to identify the proteins
- What parameters were used by the search engine (or engines)
- What criteria were used for protein identification.

Following this is a narrative description of the same information. This can be used as a rough draft for the methods section of any journal article. Although the User will undoubtedly want to clean up this computer generated text to improve its readability, it gives a place to start.

Exporting MCP supplemental table, the second step in the MCP Submission checklist, exports the [Publication report](#). In the MCP Submission procedure the User must finish step 1, describing the experimental methods, before he/she can export the Publication Report in order to ensure that the report is complete.

For publication in Proteomics journals, the User might also use the [Protein report](#) and [Peptide report](#) as supplemental supporting data.

Figure 11-9: Publication report example

```
tutorial_2_multiple_cat, Publication report created on 01/09/2014
Experiment: tutorial_2_multiple_cat
  Peak List Generator: unknown
    Version: unknown
    Charge States Calculated: unknown
    Deisotoped: unknown
    Textual Annotation: unknown
  Database Set: 2 Databases
    Database Name: a subset of the control_sprot database
      Version: unknown
      Taxonomy: All Entries
      Number of Proteins: 766
    Database Name: the control_sprot_1 database
      Version: unknown
      Taxonomy: All Entries
      Number of Proteins: 127876
  Explain Database w/ < 1000 entries:
  Does database contain common contaminants?: unknown
Search Engine Set: 2 Search Engines
  Search Engine: Mascot
    Version: 2.4.0
    Samples: All Samples
    Fragment Tolerance: 0.50 Da (Monoisotopic)
    Parent Tolerance: 1.2 Da (Monoisotopic)
    Fixed Modifications: +57 on C (Carbamidomethyl)
    Variable Modifications: +16 on M (Oxidation), +43 on n (Carbamyl)
    Database: the control_sprot_1 database (unknown version, 127876 entries)
    Digestion Enzyme: Trypsin
    Max Missed Cleavages: 1
    Probability Model:
      control_071904_01 (F001807): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_02 (F001808): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_03 (F001809): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_04 (F001810): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_05 (F001811): LFDR Model, Classifier data: Bayes, Good (f
  Search Engine: X! Tandem
    Version: CYCLONE (2010.12.01.1)
    Samples: All Samples
    Fragment Tolerance: 0.50 Da (Monoisotopic)
    Parent Tolerance: 1.2 Da (Monoisotopic)
    Fixed Modifications: +57 on C (Carbamidomethyl)
    Variable Modifications: -18 on n (Glu>pyro-Glu), -17 on n (Ammonia-loss), -17 on n (
    Database: a subset of the control_sprot database
    Digestion Enzyme: Trypsin
    Max Missed Cleavages: 2
    Probability Model:
      control_071904_01 (F001807): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_02 (F001808): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_03 (F001809): LFDR Model, Classifier data: Bayes, Good (f
      control_071904_04 (F001810): LFDR Model, No Classifier [all charge state
      control_071904_05 (F001811): LFDR Model, No Classifier [all charge state
Scaffold: Version: Scaffold_4.2.1
  Modification Metadata Set: 1541 modifications
    Source: C:\Program Files\Scaffold 4-820\parameters\unimod.xml
    Comment:
  Protein Grouping Strategy: Experiment-wide grouping with protein cluster analysis
  Peptide Threshold: 95.0% minimum
```

Samples report

The Samples report mimics the Samples View. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#).

Subsequently each row in the report represents a protein in the samples list. The number of proteins displayed depends on the current filter and threshold settings.

If **Edit > Show GO Annotations** is selected, the Go annotation information appearing in the Samples View will also be included in the Samples Report.

There are three slightly different version of this report:

- *Samples report (regular)*- See [Figure 11-10](#).
- *Samples report with clusters* - Available when protein cluster analysis is selected. It adds clusters to the regular report
- *Samples report with Isoforms* - Includes expanded protein groups to the regular report.
- *Spectrum Counting report* - It is like the regular Samples report but reports the samples quantitative values and the proteins identification probability.

Figure 11-10: Samples report columns

#
Visible?
Sample?
Identified Proteins (32)
Accession Number
Molecular Weight
Protein Grouping Ambiguity
Quantitative Variance
Taxonomy
biological adhesion
biological regulation
cell killing
cellular process
developmental process
establishment of localization
growth
immune system process
localization
locomotion
metabolic process
multi-organism process
multi-cellular organismal process
pigmentation
reproduction
reproductive process
response to stimulus
rhythmic process
viral reproduction
Gel apparatus
cytoplasm
cytoskeleton
endoplasmic reticulum
endosome
extracellular region
intracellular organelle
membrane
mitochondrion
nucleus
organelle membrane
organelle part
plasma membrane
ribosome
antioxidant activity
auxiliary transport protein activity
binding
catalytic activity
chaperone regulator activity
chemottractant activity
chemorepellent activity
electron carrier activity
enzyme regulator activity
metallochaperone activity
molecular function
molecular transducer activity
motor activity
nutrient reservoir activity
protein tag
structural molecule activity
transcription regulator activity
translation regulator activity
transporter activity
L_M
L_G

Spectrum report

The Spectrum report details all the spectra passing the current filter and threshold settings. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#). Afterwards, each entry represents a spectrum matching a peptide.

Figure 11-11: Spectrum report columns

Experiment name
Biological sample category
Biological sample name
MS/MS sample name
Protein name
Protein accession numbers
Database sources
Protein molecular weight (Da)
Protein identification probability
Exclusive unique peptide count
Exclusive unique spectrum count
Total spectrum count
Percentage of total spectra
Percentage sequence coverage
Manual validation
Assigned
Spectrum name
Peptide sequence
Previous amino acid
Next amino acid
Peptide identification probability
SEQUEST XCorr score
SEQUEST Dcn score
X! Tandem -log(e) score
Number of enzymatic termini
Fixed modifications identified by spectrum
Variable modifications identified by spectrum
Observed m/z
Actual peptide mass (AMU)
Calculated + 1H Peptide Mass (AMU)
Spectrum change
Actual minus calculated peptide mass (AMU)
Actual minus calculated peptide mass (PPM)
Total Ion Current
Peptide start index
Peptide stop index
Exclusive
Other Proteins

Column quick notes:

- The first 14 columns of the table provide information available in the Protein Report identifying the sample and the protein.
- The Manual validation column reports if the User manually validated a spectrum. This is done by selecting or deselecting the check-box in the “Valid” column shown in the

spectrum table displayed in the peptides pane. One of the following possible statuses of the Valid check box can appear in the report:

- *Possibly Correct* - The User accepted the status of the box resulting from Scaffold analysis and did not touch it.
- *Correct* - The User deselected and then selected the box again.
- *Unchecked box* - If the box remains unchecked the spectrum does not appear in this report.
- Number of enzymatic termini (NTT). When the digestion enzyme is trypsin, this tells if the peptide is tryptic (2) semi-tryptic (1) or non-tryptic (0).

Peptide report

The Peptide report details all the peptides that pass the current filter and thresholds settings. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#).

Afterwards, every row represents a peptide in each of the samples present in the Scaffold experiment. For example, if there are 3 samples each with 100 peptides, there will be 300 rows in the report. Even if several spectra match a peptide, the peptide only gets one line in this report.

Figure 11-12: Peptide report columns

Experiment name
Biological sample category
Biological sample name
MS/MS sample name
Protein name
Protein accession numbers
Database sources
Protein molecular weight (Da)
Protein identification probability
Exclusive unique peptide count
Exclusive unique spectrum count
Total spectrum count
Percentage of total spectra
Percentage sequence coverage
Peptide sequence
Previous amino acid
Next amino acid
Best Peptide identification probability
Best SEQUEST Xcorr score
Best SEQUEST Dcn score
Best X! Tandem -log(e) score
Number of identified +1H spectra
Number of identified +2H spectra
Number of identified +3H spectra
Number of identified +4H spectra
Number of enzymatic termini
Calculated +1H Peptide Mass (AMU)
Median Retention Time
Total Precursor Intensity
Total TIC
Peptide start index
Peptide stop index
Star Category
Assigned
Other Proteins

Columns quick notes:

- The first 14 columns repeat the information available in the Protein Report which identify the sample and the protein.
- Next comes the peptide sequence followed by the best scores for the spectrum matching it. Then there are columns showing how many spectra matched the peptides in each charge state and a column for the number of tryptic termini (NTT)

There are two different versions of this report:

- *Peptide Report (regular)* - See [Figure 11-12](#)
- *Peptide Quantitative report* - which exports similar information as the regular report does but organizes it emphasizing the various quantitative values available in the experiment for each peptide in every sample, see [Figure 11-13](#)

Figure 11-13: Peptide quantitative report columns

	Total Spectrum Count				Weighted Spectrum Count				Total Precursor Intensity				Total TIC				Median Retention Time			
	1	M	1	G	1	G	1	G	1	M	1	G	1	M	1	G	1	M	1	G
Protein																				
Peptide																				
Mudpit_bovine_mudpit_10																				
bovine_mudpit_10																				
bovine_mudpit_11																				
bovine_mudpit_12																				
bovine_mudpit_13																				

Protein report

The protein report details all the proteins passing the current filter and threshold settings. This report is designed to be used as part of the supplemental information supporting a journal article. The report header rows identify the data and how it was created, which is the same information that is contained in the [Publication report](#). Then there is a single report entry for each protein in every sample. For example, if there are 3 samples each with the same 12 proteins, there will be 36 rows in this report.

Figure 11-14: Protein report columns

Experiment name	Biological sample category	Biological sample name	MS/MS sample name	Protein name	Protein accession numbers	Database sources	Protein molecular weight (Da)	Protein identification probability	Exclusive unique peptide count	Exclusive unique spectrum count	Total spectrum count	Percentage of total spectra	Percentage sequence coverage	Peptide sequence	Previous amino acid	Next amino acid	Best Peptide identification probability	Best SEQUEST XCorr score	Best SEQUEST DCn score	Best X! Tandem -log(e) score	Number of identified +1H spectra	Number of identified +2H spectra	Number of identified +3H spectra	Number of identified +4H spectra	Number of enzymatic termini	Calculated +1H Peptide Mass (AMU)	Median Retention Time	Total Precursor Intensity	Total TIC	Peptide start index	Peptide stop index	Star Category	Assigned	Other Proteins
-----------------	----------------------------	------------------------	-------------------	--------------	---------------------------	------------------	-------------------------------	------------------------------------	--------------------------------	---------------------------------	----------------------	-----------------------------	------------------------------	------------------	---------------------	-----------------	---	--------------------------	------------------------	------------------------------	----------------------------------	----------------------------------	----------------------------------	----------------------------------	-----------------------------	-----------------------------------	-----------------------	---------------------------	-----------	---------------------	--------------------	---------------	----------	----------------

Columns quick notes:

- The first 4 columns identify the experiment, the biological sample, its category and the MS/MS sample (or run).
- This is followed by columns that identify the protein by name, accession number, database where the accession resides, and the protein's mass.
- The remainder of the columns provide the results of the analysis.

There are three different versions of this report:

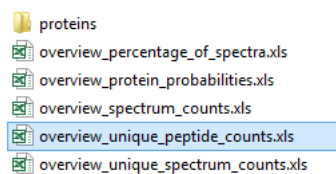
- Protein report (regular)* - See [Figure 11-14](#)
- Protein cluster report* - Available when protein cluster analysis is selected. It adds clusters to the regular report
- Protein Accession Number Report* - Similar to the regular report, its purpose is to provide an easy way to look up the protein description for each accession number. The report also provides the name of the database that was used for searching the data.

Current View report

The Current View report contains the information that is displayed in the current view. This report is applicable for the [Samples View](#), the [Proteins View](#), and the [Publish View](#).

Complete report

This export is meant to provide the full results of the current analysis in a series of *.XML files saved in a separate directory. The directory created contains the following list of files:



Chapter 11

Reports

Appendix

- [A.Algorithms References](#)
- [B.Terminology](#)
- [C.Terminology comparison between Scaffold 4 and Scaffold 3](#)

A. Algorithms References

The algorithm for calculating the peptide probabilities from the search engine scores is described in:

Keller (2002)

Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold R., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.*

Anal. Chem., 2002, 74 (20), pp 5383–5392 DOI: 10.1021/ac025747h

The algorithm for calculating the protein probabilities from the peptide probabilities is described in:

Nesvizhskii (2003)

Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., *A statistical model for identifying proteins by tandem mass spectrometry.*

Anal. Chem., 2003, 75 (17), pp 4646–4658 DOI: 10.1021/ac0341261

The algorithm for combining results from multiple searches is described in:

Searle (2008)

Searle, B.C., Turner, M., Nesvizhskii, A.I., *Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies.*

J Proteome Res., 2008, 7(1), pp 245-53 DOI: 10.1021/pr070540w

The algorithm for grouping proteins across samples is described in

Searle (2010)

Searle, B.C., *Scaffold: A bioinformatic tool for validating MS/MS-based proteomics studies*

Proteomics, 2010, 10(6), pp 1265-9. DOI: 10.1002/pmic.200900437.

The algorithm for X! Tandem is described in:

Craig (2003)

Craig, R., Beavis, R.C., *A method for reducing the time required to match protein sequences with tandem mass spectra.*

Rapid Communications in Mass Spectrometry, 2003, 17(20), pp 2310-6. DOI:10.1002/rcm.1198

Purity corrections calculations for iTRAQ data:

Shadforth (2005)

Shadforth, I.P., Dunkley T.P., Lilley K.S. and Bessant C., *i-Tracker: For quantitative proteomics using iTRAQ™*

BMC Genomics, 2005, 6:145. DOI:10.1186/1471-2164-6-145

Pavelka (2008)

Pavelka, N., Fournier, M.,L., Swanson, S.,K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L. and Washburn M.,P., *Statistical Similarities between Transcriptomics and Quantitative Shotgun Proteomics Data*

Mol Cell Proteomics April, 2008 7: 631-644. DOI:10.1074/mcp.M700240-MCP200

Reference for calculating emPAI:

Ishihama (2005)

Ishihama, Y.,Oda, Y., Tabata, T., Sato,T., Nagasu, T., Rappsilber, J., Mann, M., *Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein*

Molecular & Cellular Proteomics, 2005, 4, 1265-1272. DOI: 10.1074/mcp.M500061-MCP200

Reference for calculating SAF:

Zhang (2010)

Zhang, Y., Wen, Z., Washburn, M.P., & Florens, L., *Refinements to Label Free Proteome Quantitation: How to Deal with Peptides Shared by Multiple Proteins*

Anal. Chem., 2010, 82(6):2272-81. DOI: 10.1021/ac9023999

References for Precursor intensity quantitation:

Bantscheff (2007).

Bantscheff M., Schirle M., Sweetman G.,Rick J. and Kuster. B. *Quantitative mass spectrometry in proteomics: a critical view*

Anal Bioanal Chem, 2007, 389:1017–1031. DOI:10.1007/s00216-007-1486-6

Lai (2013)

Xianyin Lai, Lianshui Wang, and Frank A. Witzmann, *Issues and Applications in Label-Free Quantitative Mass Spectrometry*

International Journal of Proteomics, 2013, vol. 2013, Article ID 756039, 13 pages.
DOI:10.1155/2013/756039

Appendix

Raubenheimer (1992)

Raubenheimer, D. and Simpson, S. L., *Analysis of covariance: an alternative to nutritional indices*.

Entomologia Experimentalis et Applicata, 1992, 62: 221–231. DOI: 10.1111/j.1570-7458.1992.tb00662.x

References for Fisher's Exact test:

Zhang (2006)

Zhang, B., VerBerkmoes, N.C., Langston, M. A., Uberbacher, E., Hettich, R. L., Samatova, N. F. *Detecting differential and correlated protein expression in label-free shotgun proteomics*

J. Proteome Res., 2006, 5 (11), pp 2909–2918. DOI: 10.1021/pr0600273

B. Terminology

BioSample

Scaffold calls BioSample a physical sample, such as a drop of blood or biopsy from a patient, or a tissue sample from a model organism or cell line. The proteins or peptides in a BioSample are typically separated by 2D gels or liquid chromatography into several spots, bands, or fractions, each of which becomes one mass spectrometry sample or MS Sample. One BioSample is therefore typically made up of several MS samples. Both BioSamples and MS Samples are often referred to by practitioners just as “samples”.

- When running Scaffold Q+ or Scaffold Q+S, quantitative multiplexed samples are initially loaded in Scaffold and referred to as BioSamples.

Exclusive Spectrum Count

Total number of spectra associated with this protein including those shared with other proteins

Mascot File Names

If a file containing Mascot data is named something like F*.DAT, for example: F123456.dat or F987654.dat, Scaffold looks inside the Mascot file to search for the name of the original file that Mascot searched. Since that name is more likely to be meaningful, Scaffold uses it to name the MS sample. On the Load Data page, Scaffold displays the “original” name followed by the F*.DAT name assigned by Mascot in parentheses. On the Samples and Proteins pages, Scaffold uses only the original name. The MS Sample name can be changed on the Samples View, see “[Sample Information Pane](#)” on page 141.

Mascot Output Files

Mascot is a search engine distributed by Matrix Science. Scaffold loads Mascot result files created with the following format: *.DAT. When Mascot is used through Proteome Discoverer the results are stored in *.MSF type of file.

NSAF

The Normalized Spectral Abundance Factor (NSAF) is a modified version of spectral counting. It was introduced and defined by the Washburn Lab group at the Stowers Institute to account for the effects of protein length, as large proteins tend to contribute to the spectral counts a greater amount of peptide/spectra than smaller ones do. For more detailed information visit the Washburn Lab website: <http://research.stowers.org/proteomics/Quant.html>.

Sequest Output Files

Sequest is one of the search engines distributed by Thermo Scientific. Depending on the platform used to run it Sequest creates files with the following format: *.DTA and *.OUT, *.MS2 and *.SQT, and *.SRF types of file. The new platform developed by Thermo Scientific called Proteome Discoverer creates output files with extension:*.MFS

TIC- Total Ion Current

Appendix

The total ion current (TIC) is the sum of the areas under all the peaks contained in a MS/MS spectrum. Scaffold assumes that the area under a peak is proportional to the height of the peak and approximates the TIC value by summing the intensity of the peaks contained in the peak list associated to a MS/MS sample.

Total Unique Peptide Count

Number of different amino acid sequences that are associated with a protein including those shared with other proteins

Total Unique Spectrum Count

Number of unique spectra associated with a protein including those shared with other proteins

C. Terminology comparison between Scaffold 4 and Scaffold 3

Starting from Scaffold version 4 Proteome Software added new terms to capture different types of evidence used in protein clustering. These new terms affect the display options in the Samples View. The tables below indicate the correspondence with Scaffold 3 display option terms (The User might want to print these tables).

Table 1: New Terminology

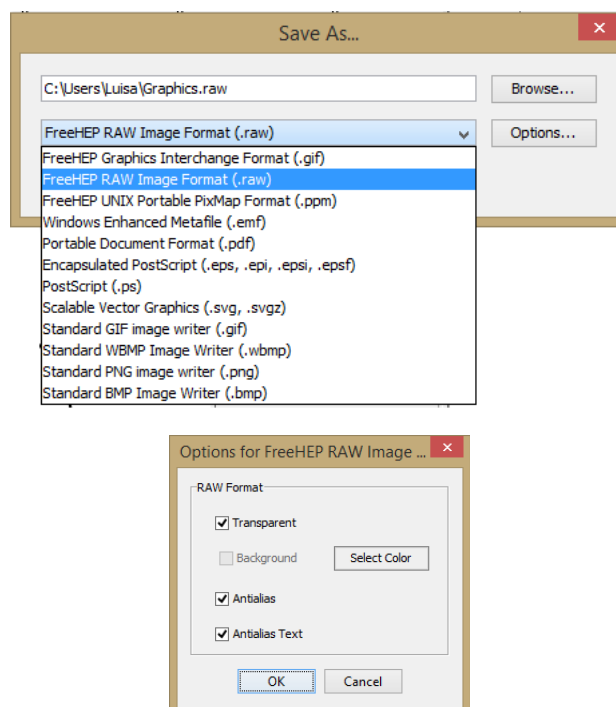
Unique peptides	Peptides with different amino acid sequences, regardless of any modifications
Unique spectra	Spectra that differ in amino acid sequence, charge state or modifications
Exclusive	Associated with a single protein group
Total	Associated with a protein group, whether or not it is shared with other protein groups

Table 2: Terminology comparison between Scaffold 4 and Scaffold 3

Scaffold 4 term	Scaffold 3 term	Description
Exclusive Unique Peptide Count	Number of Unique Peptides	Number of different amino acid sequences that are associated only with this protein
Exclusive Unique Spectrum Count	Number of Unique Spectra	Number of unique spectra associated only with this protein
Exclusive Spectrum Count	Number of Assigned Spectra	Number of spectra associated only with this protein
Total Spectrum Count	Unweighted Spectrum Count	Total number of spectra associated with this protein including those shared with other proteins
Total Unique Peptide Count	N/A	Number of different amino acid sequences that are associated with this protein including those shared with other proteins
Total Unique Spectrum Count	N/A	Number of unique spectra associated with this protein including those shared with other proteins

D. Mouse Right Click Commands

- **BLAST Peptide Sequence** - opens <http://www.ncbi.nlm.nih.gov/blast/> site for current peptide
- **BLAST protein sequence** - opens <http://www.ncbi.nlm.nih.gov/blast/> site for current protein.
- **Copy All Data** - copies all the data listed in the table shown in the current pane to the clipboard.
- **Copy Image** - copies the image of the current view and current pane to the clipboard.
- **Copy Peak List** - copies the pick list of the current spectrum to the clipboard.
- **Copy Protein Sequence** - copies the sequence to the clipboard.
- **Copy publication Sized JPEG** - for publication purposes.
- **Copy Selected Cell** - from the table copies selected cell to the clipboard.
- **Copy Selected row** - from the table copies selected row to the clipboard.
- **Copy WMF/EMF** - copies picture using Windows Meta-file formats which are portable between applications. They contain both vector graphics and bitmap components. Images can be edited and scaled without compromising their resolution.
- **Delete Biological Sample** - a window pops up asking to confirm deletion.
- **Display parent Ions** - toggle function.
- **Display unknown markers** - toggle function
- **Export to Excel** - export information in current tab table
- **Edit BioSample** - See [Edit BioSample](#).
- **Print** - print image of current view and pane.
- **Queue Files for Loading** - See [Queue Files for Loading](#).
- **Save as** - provides the option of saving pictures in a large variety of graphical formats.



- **Save JPEG Image** - saves image of current view and pane JPEG format.
- **Show Fixed Modifications** - it toggles the function of highlighting fixed modifications along the sequence.
- **Use Amino Acid Finder** - it toggles the activation of the tool tip that shows the peptides along the sequence.
- **Use Peak-finder** - displays the tool tip for the different peaks if checked.
- **Use PPM Masses** - toggle function.
- **Zoom Out** - zoom function.

Appendix

Index

A

Algorithms references.....	206
Appendix.....	205
References algorithms.....	206
Assumptions for the manual.....	8

C

Configure peptide threshold dialog box.....	131
Conventions used in the manual ...	7
Copyright	3

D

Display pane in the Samples View ..	135
display options	135
Req Mods	136
search feature.....	137

E

Export	
mzIdentML	191
Pride	192
ProtXML.....	191
Scaffold perSPECTives	192
ScaffoldPTM	192
Spectra	190
Subset database.....	189, 190
Export reports	189

F

FASTA databases	68
FASTA databases in Scaffold.....	53

FDR	
How Scaffold calculates.....	134
FDR filtering	
Filtering Samples	134
Filtering samples.....	129
Custom peptide filters	131
FDR filtering	134
Minimum number of peptides..	130
Peptide threshold	130
Protein thresholds	129

G

Gene Ontology terms pane.....	148
GO bar charts	149
GO Pie charts	148

L

LFDR-based scoring system	26
licensing for Scaffold.....	13
Loading Wizard.....	38

M

Mouse right click commands	212
mzIdentML	
export.....	191

O

Organization of the manual.....	8
---------------------------------	---

P

Peptide report	201
PeptideProphet	27
Precursor intensity quantitation .	179
Calculations	180
Mascot Distiller.....	182
MaxQuant	182

Performing quantitation	183
Preparing data for	181
Proteome Discoverer	182
Spectrum Mill	182

Pride	
export.....	192
Protein and PeptideProphet.....	26
Protein annotation preferences..	126
Protein Information pane in the Samples View	139
Protein list	124
Hidden proteins	128
Protein cluster	125
Protein group.....	124
Proteins of interest	127
Sorting Feature	127
ProteinProphet.....	27
proteins	
hiding in the Samples View	128
proteins of interest	
identifying in the Samples View starring proteins	127
ProtXML export.....	191
ProtXML report.....	191
Publication report.....	198

Q

Quantify View	
Gene Ontology terms pane.....	148
GO bar charts.....	149
GO pie charts.....	148
Quantitation in Scaffold	
Precursor intensity	183

R

Release Information.....	3
Renewing time based license key	17
Reports	189

reports	
Peptide.....	201
ProtXML.....	191
Publication	198
Samples.....	199

S

Samples report	199
Samples Table.....	123
Samples View	
advanced Searchconfigure	
advanced protein filter.....	138
Display pane	135
hiding proteins from	128
Protein Information pane.....	139
proteins of interest	127
sorting feature	127
Scaffold	
tiered licensing for.....	13
Scaffold 3	
terminnology comparison Scaffold	
4	211
Scaffold perSPECTives	
export.....	192
ScaffoldPTM	
export.....	192
Scoring algorithm	
LFDR-based scoring	127
PeptideProphet scoring.....	127
Scoring system	
LFDR-based.....	26
PeptideProphet	27
ProteinProphet	27
show hidden proteins	128
Skyline	194
sorting feature	
Samples View	127
Special information about the	
manual	7

T

Terminology	209
BioSample.....	209
Terminology comparison between	
Scaffold 3 and Scaffold 4	211

U

Upgrading Scaffold to Scaffold Q+ or	
Scaffold Q+S.....	16
Using the manual	7